

PATENT COOPERATION TREATY

PCT

REC'D 16 OCT 2001
WIPO PCT

INTERNATIONAL PRELIMINARY EXAMINATION REPORT

(PCT Article 36 and Rule 70)

Applicant's or agent's file reference PHM70564 PCT	FOR FURTHER ACTION		See Notification of Transmittal of International Preliminary Examination Report (Form PCT/IPEA/416)
International application No. PCT/US00/20401	International filing date (day/month/year) 27/07/2000	Priority date (day/month/year) 27/07/1999	
International Patent Classification (IPC) or national classification and IPC G06K9/00			
<p>Applicant ZENECA LIMITED et al.</p> <p>1. This international preliminary examination report has been prepared by this International Preliminary Examining Authority and is transmitted to the applicant according to Article 36.</p> <p>2. This REPORT consists of a total of 6 sheets, including this cover sheet.</p> <p><input type="checkbox"/> This report is also accompanied by ANNEXES, i.e. sheets of the description, claims and/or drawings which have been amended and are the basis for this report and/or sheets containing rectifications made before this Authority (see Rule 70.16 and Section 607 of the Administrative Instructions under the PCT).</p> <p>These annexes consist of a total of sheets.</p>			
<p>3. This report contains indications relating to the following items:</p> <ul style="list-style-type: none"> I <input checked="" type="checkbox"/> Basis of the report II <input type="checkbox"/> Priority III <input type="checkbox"/> Non-establishment of opinion with regard to novelty, inventive step and industrial applicability IV <input type="checkbox"/> Lack of unity of invention V <input checked="" type="checkbox"/> Reasoned statement under Article 35(2) with regard to novelty, inventive step or industrial applicability; citations and explanations supporting such statement VI <input type="checkbox"/> Certain documents cited VII <input type="checkbox"/> Certain defects in the international application VIII <input checked="" type="checkbox"/> Certain observations on the international application 			

Date of submission of the demand 27/02/2001	Date of completion of this report 12.10.2001
Name and mailing address of the international preliminary examining authority: European Patent Office D-80298 Munich Tel. +49 89 2399 - 0 Tx: 523656 epmu d Fax: +49 89 2399 - 4465	Authorized officer Kessler, C Telephone No. +49 89 2399 2582



**INTERNATIONAL PRELIMINARY
EXAMINATION REPORT**

International application No. PCT/US00/20401

I. Basis of the report

1. With regard to the **elements** of the international application (*Replacement sheets which have been furnished to the receiving Office in response to an invitation under Article 14 are referred to in this report as "originally filed" and are not annexed to this report since they do not contain amendments (Rules 70.16 and 70.17)*):
Description, pages:

1-72 as originally filed

Claims, No.:

1-48 as originally filed

Drawings, sheets:

1-64 as originally filed

2. With regard to the **language**, all the elements marked above were available or furnished to this Authority in the language in which the international application was filed, unless otherwise indicated under this item.

These elements were available or furnished to this Authority in the following language: , which is:

- the language of a translation furnished for the purposes of the international search (under Rule 23.1(b)).
- the language of publication of the international application (under Rule 48.3(b)).
- the language of a translation furnished for the purposes of international preliminary examination (under Rule 55.2 and/or 55.3).

3. With regard to any **nucleotide and/or amino acid sequence** disclosed in the international application, the international preliminary examination was carried out on the basis of the sequence listing:

- contained in the international application in written form.
- filed together with the international application in computer readable form.
- furnished subsequently to this Authority in written form.
- furnished subsequently to this Authority in computer readable form.
- The statement that the subsequently furnished written sequence listing does not go beyond the disclosure in the international application as filed has been furnished.
- The statement that the information recorded in computer readable form is identical to the written sequence listing has been furnished.

4. The amendments have resulted in the cancellation of:

- the description, pages:
- the claims, Nos.:

INTERNATIONAL PRELIMINARY
EXAMINATION REPORT

International application No. PCT/US00/20401

the drawings, sheets:

5. This report has been established as if (some of) the amendments had not been made, since they have been considered to go beyond the disclosure as filed (Rule 70.2(c)): *(Any replacement sheet containing such amendments must be referred to under item 1 and annexed to this report.)*

6. Additional observations, if necessary:

**V. Reasoned statement under Article 35(2) with regard to novelty, inventive step or industrial applicability;
citations and explanations supporting such statement**

1. Statement

Novelty (N)	Yes:	Claims 6, 8, 11-16, 22, 24, 27-32, 38, 40, 44-48
	No:	Claims 1-5, 7, 10, 17-21, 23, 26, 33-37, 39, 42
Inventive step (IS)	Yes:	Claims 13-16, 29-32, 45-48
	No:	Claims 6, 8, 9, 11, 12, 22, 24, 27, 28, 38, 40, 43, 44
Industrial applicability (IA)	Yes:	Claims 1 - 48
	No:	Claims

2. Citations and explanations
see separate sheet

VIII. Certain observations on the international application

The following observations on the clarity of the claims, description, and drawings or on the question whether the claims are fully supported by the description, are made:
see separate sheet

INTERNATIONAL PRELIMINARY
EXAMINATION REPORT - SEPARATE SHEET

International application No. PCT/US00/20401

1. D1 to D3 are referred to as the documents cited in the International Search Report, according to the sequence in which they are listed there.
NB: For D2, cf also <http://genome-www.stanford.edu/clustering/> for enhanced graphics.

2. Claims 1-16, 17-32 and 33-48 define in completely parallel manner methods, systems and computer-readable memory media. In the following reference is made representatively only to claims 1-16.

Re Item V
Reasoned statement under Rule 66.2(a)(ii) with regard to novelty, inventive step or industrial applicability; citations and explanations supporting such statement

3. The invention relates to data mining in large datasets, and in particular to clustering the raw data in order to find groups of (in some way) similar data entries. This is a notorious problem wherever large datasets having unknown characteristics are concerned.
4. D2 relates to clustering gene expression measurements. It anticipates the subject matter of claims 1 to 5, 7 and 10:
 - 4.1 Claim 1: D2 discloses a method of operating on data, the method comprising: providing at least one user-defined grouping rule (the user being the authors of the paper; the grouping depending on the fluorescence ratio) for grouping the data into a user-definable number of groups (similar; 3 groups); and applying at least one of the grouping rules to the data. Cf the abstract and page 14863, right column, third paragraph ("Although various ..."), as well as page 14864, right column, the last paragraph before "RESULTS" ("Display.")
 - 4.2 Claim 2: D2's data are provided in a table (cf page 14863, right column, the phrase in the middle of the second paragraph: "We aim to use these methods ..."), and applies the grouping rule to one (author-selected) column of the table (fluorescence intensity ratios, cf page 14864, left column, the last phrase of the first paragraph, and the second paragraph).

INTERNATIONAL PRELIMINARY
EXAMINATION REPORT - SEPARATE SHEET

International application No. PCT/US00/20401

4.3 Claim 3: the breakpoints are defined by the values of the fluorescence (log) ratios, cf. again page 14864, right column, the last paragraph before "RESULTS".

4.4 Claim 4: D2 presents the grouped data in a manner that visually distinguishes the groups, cf again the same paragraph.

4.5 Claim 5: in D2, an aspect of the data is coloured according to the rules (black, red, green; cf again said paragraph).

4.6 Claim 7: the fluorescence log ratio is a numeric value.

4.7 Claim 10 is trivial (in the mathematical sense).

5. Furthermore, the subject matter of claims 6, 8, 9, 11 and 12 is obvious to the skilled person:

5.1 Claim 6: operating on tables within spreadsheets is notorious. The skilled person would likely place the table of D2 into a spreadsheet. (Automatically) colouring elements of such spreadsheet tables is therefore obvious in view of D2.

5.2 Claim 7, the "textual values" alternative: sorting table entries alphabetically according to one column is notorious in spreadsheet applications.

5.3 Automatically determining breakpoints, ie the partitioning between clusters, is one of the main features of cluster analysis, cf D3, page 21, right column, the paragraph on "Cluster Analysis". D2 makes use of unsupervised clustering to generate its dendograms, cf page 14863, right column, the middle of the second paragraph ("Clustering methods ...") until the end of the column.

5.4 Claim 9: colouring the cell entry or the background of the cell represent obvious alternatives, from which the skilled person will choose according to the circumstances.

5.5 Claims 11 and 12 are confusing in that they define the scoring step after the grouping step. Usually, scoring takes place in order to define the grouping (cf D2,

**INTERNATIONAL PRELIMINARY
EXAMINATION REPORT - SEPARATE SHEET**

International application No. PCT/US00/20401

page 14864, left column, paragraph 3, "Metrics"). However, the claims can be interpreted as colouring and sorting the data values as known from D2 (generation of the spanning tree, cf page 14863, right column, last paragraph, to the end of "Ordering of Data Tables" on page 12864, right column).

Re Item VIII

Certain observations on the international application

- 6.1 Claims 11 and 12 are ambiguous and hence indefinite as explained above under item 5.5.
- 6.2 Similarly, claim 13 does not define the scoring properly. It is not clear what is the motivation behind the scoring, nor what the parameters might be. It appears that this claim relates closely to an embodiment of the present application (eg comparing columns), and should hence have been formulated in more specific terms.

PATENT COOPERATION TREATY

PCT

INTERNATIONAL SEARCH REPORT

(PCT Article 18 and Rules 43 and 44)

Applicant's or agent's file reference PHM70564 PCT	FOR FURTHER ACTION see Notification of Transmittal of International Search Report (Form PCT/ISA/220) as well as, where applicable, item 5 below.	
International application No. PCT/US 00/ 20401	International filing date (<i>day/month/year</i>) 27/07/2000	(Earliest) Priority Date (<i>day/month/year</i>) 27/07/1999
Applicant ZENECA LIMITED		

This International Search Report has been prepared by this International Searching Authority and is transmitted to the applicant according to Article 18. A copy is being transmitted to the International Bureau.

This International Search Report consists of a total of 3 sheets.

It is also accompanied by a copy of each prior art document cited in this report.

1. Basis of the report

a. With regard to the **language**, the international search was carried out on the basis of the international application in the language in which it was filed, unless otherwise indicated under this item.

the international search was carried out on the basis of a translation of the international application furnished to this Authority (Rule 23.1(b)).

b. With regard to any **nucleotide and/or amino acid sequence** disclosed in the international application, the international search was carried out on the basis of the sequence listing :

contained in the international application in written form.

filed together with the international application in computer readable form.

furnished subsequently to this Authority in written form.

furnished subsequently to this Authority in computer readable form.

the statement that the subsequently furnished written sequence listing does not go beyond the disclosure in the international application as filed has been furnished.

the statement that the information recorded in computer readable form is identical to the written sequence listing has been furnished

2. **Certain claims were found unsearchable** (See Box I).

3. **Unity of invention is lacking** (see Box II).

4. With regard to the **title**,

the text is approved as submitted by the applicant.

the text has been established by this Authority to read as follows:

5. With regard to the **abstract**,

the text is approved as submitted by the applicant.

the text has been established, according to Rule 38.2(b), by this Authority as it appears in Box III. The applicant may, within one month from the date of mailing of this international search report, submit comments to this Authority.

6. The figure of the **drawings** to be published with the abstract is Figure No.

as suggested by the applicant.

because the applicant failed to suggest a figure.

because this figure better characterizes the invention.

2

None of the figures.

INTERNATIONAL SEARCH REPORT

International Application No
PCT/US 00/20401A. CLASSIFICATION OF SUBJECT MATTER
IPC 7 G06K9/00

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)
IPC 7 G06K

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practical, search terms used)
EPO-Internal, COMPENDEX, INSPEC, WPI Data, IBM-TDB, PAJ

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	ANONYMOUS: "Dynamic Layout Mechanism for the Massive-Node Server Status Monitor" IBM TECHNICAL DISCLOSURE BULLETIN, vol. 36, no. 5, 1 May 1993 (1993-05-01), pages 169-170, XP000408951 New York, US the whole document ---	1-48 -/-

 Further documents are listed in the continuation of box C. Patent family members are listed in annex.

° Special categories of cited documents :

- "A" document defining the general state of the art which is not considered to be of particular relevance
- "E" earlier document but published on or after the international filing date
- "L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)
- "O" document referring to an oral disclosure, use, exhibition or other means
- "P" document published prior to the international filing date but later than the priority date claimed

- "T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention
- "X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone
- "Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art.
- "&" document member of the same patent family

Date of the actual completion of the international search

Date of mailing of the international search report

12 January 2001

22/01/2001

Name and mailing address of the ISA
European Patent Office, P.B. 5818 Patentlaan 2
NL - 2280 HV Rijswijk
Tel. (+31-70) 340-2040, Tx. 31 651 epo nl,
Fax: (+31-70) 340-3016

Authorized officer

Granger, B

INTERNATIONAL SEARCH REPORT

International Application No

PCT/US 00/20401

C.(Continuation) DOCUMENTS CONSIDERED TO BE RELEVANT

Category	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A ✓	<p>EISEN M B ET AL: "Cluster analysis and display of genome-wide expression patterns" PROCEEDINGS OF THE NATIONAL ACADEMY OF SCIENCES OF USA, US, NATIONAL ACADEMY OF SCIENCE, WASHINGTON, VOL. 95, December 1998 (1998-12), pages 14863-14868, XP002140966 ISSN: 0027-8424 page 14863, right-hand column, paragraph 3; figure 2</p> <p>---</p>	1-48
A ✓	<p>STANTON D T ET AL: "Application of nearest-neighbor and cluster analyses in pharmaceutical lead discovery" JOURNAL OF CHEMICAL INFORMATION AND COMPUTER SCIENCES, JAN.-FEB. 1999, ACS, USA, VOL. 39, NO. 1, pages 21-27, XP000971515 ISSN: 0095-2338 the whole document</p> <p>-----</p>	1-48

PATENT COOPERATION TREATY

From the INTERNATIONAL BUREAU

PCT

NOTIFICATION OF ELECTION
(PCT Rule 61.2)

Date of mailing (day/month/year) 30 April 2001 (30.04.01)	To: Commissioner US Department of Commerce United States Patent and Trademark Office, PCT 2011 South Clark Place Room CP2/5C24 Arlington, VA 22202 ETATS-UNIS D'AMERIQUE in its capacity as elected Office
International application No. PCT/US00/20401	Applicant's or agent's file reference PHM70564 PCT
International filing date (day/month/year) 27 July 2000 (27.07.00)	Priority date (day/month/year) 27 July 1999 (27.07.99)
Applicant LERMAN, Charles, L.	

1. The designated Office is hereby notified of its election made:

in the demand filed with the International Preliminary Examining Authority on:

27 February 2001 (27.02.01)

in a notice effecting later election filed with the International Bureau on:

2. The election was

was not

made before the expiration of 19 months from the priority date or, where Rule 32 applies, within the time limit under Rule 32.2(b).

The International Bureau of WIPO 34, chemin des Colombettes 1211 Geneva 20, Switzerland Facsimile No.: (41-22) 740.14.35	Authorized officer Céline Faust Telephone No.: (41-22) 338.83.38
---	--

PATENT COOPERATION TREATY

PCT

NOTIFICATION OF THE RECORDING
OF A CHANGE(PCT Rule 92bis.1 and
Administrative Instructions, Section 422)

Date of mailing (day/month/year) 20 September 2001 (20.09.01)	
Applicant's or agent's file reference PHM70564 PCT	IMPORTANT NOTIFICATION
International application No. PCT/US00/20401	International filing date (day/month/year) 27 July 2000 (27.07.00)

From the INTERNATIONAL BUREAU

To:

BIRD, Donald, J.
Pillsbury Winthrop LLP
1600 Tysons Boulevard
McLean, VA 22102
ETATS-UNIS D'AMERIQUE

1. The following indications appeared on record concerning:				
<input type="checkbox"/> the applicant <input type="checkbox"/> the inventor <input checked="" type="checkbox"/> the agent <input type="checkbox"/> the common representative				
Name and Address BIRD, Donald, J. Pillsbury Winthrop LLP 1100 New York Avenue, N.W. Washington, DC 20005 United States of America	State of Nationality		State of Residence	
	Telephone No.		(202) 861-3000	
	Facsimile No.		(202) 822-0944	
	Teleprinter No.			
2. The International Bureau hereby notifies the applicant that the following change has been recorded concerning:				
<input type="checkbox"/> the person <input type="checkbox"/> the name <input checked="" type="checkbox"/> the address <input type="checkbox"/> the nationality <input type="checkbox"/> the residence				
Name and Address BIRD, Donald, J. Pillsbury Winthrop LLP 1600 Tysons Boulevard McLean, VA 22102 United States of America	State of Nationality		State of Residence	
	Telephone No.		(703) 905-2000	
	Facsimile No.		(703) 905-2500	
	Teleprinter No.			
3. Further observations, if necessary:				
4. A copy of this notification has been sent to:				
<input checked="" type="checkbox"/> the receiving Office <input type="checkbox"/> the International Searching Authority <input checked="" type="checkbox"/> the International Preliminary Examining Authority		<input type="checkbox"/> the designated Offices concerned <input checked="" type="checkbox"/> the elected Offices concerned <input type="checkbox"/> other:		

The International Bureau of WIPO 34, chemin des Colombettes 1211 Geneva 20, Switzerland Facsimile No.: (41-22) 740.14.35	Authorized officer François BAECHLER Telephone No.: (41-22) 338.83.38
---	---

PATENT COOPERATION TREATY

PCT

NOTIFICATION OF THE RECORDING
OF A CHANGE(PCT Rule 92bis.1 and
Administrative Instructions, Section 422)

From the INTERNATIONAL BUREAU

To:

BIRD, Donald, J.
Pillsbury Winthrop LLP
1600 Tysons Boulevard
McLean, VA 22102
ETATS-UNIS D'AMERIQUE

Date of mailing (day/month/year) 20 September 2001 (20.09.01)	
Applicant's or agent's file reference PHM70564 PCT	IMPORTANT NOTIFICATION
International application No. PCT/US00/20401	International filing date (day/month/year) 27 July 2000 (27.07.00)

1. The following indications appeared on record concerning:

the applicant the inventor the agent the common representative

Name and Address BIRD, Donald, J. Pillsbury Winthrop LLP 1100 New York Avenue, N.W. Washington, DC 20005 United States of America	State of Nationality	State of Residence
	Telephone No. (202) 861-3000	
	Facsimile No. (202) 822-0944	
	Teleprinter No.	

2. The International Bureau hereby notifies the applicant that the following change has been recorded concerning:

the person the name the address the nationality the residence

Name and Address BIRD, Donald, J. Pillsbury Winthrop LLP 1600 Tysons Boulevard McLean, VA 22102 United States of America	State of Nationality	State of Residence
	Telephone No. (703) 905-2000	
	Facsimile No. (703) 905-2500	
	Teleprinter No.	

3. Further observations, if necessary:

4. A copy of this notification has been sent to:	
<input checked="" type="checkbox"/> the receiving Office	<input type="checkbox"/> the designated Offices concerned
<input type="checkbox"/> the International Searching Authority	<input checked="" type="checkbox"/> the elected Offices concerned
<input checked="" type="checkbox"/> the International Preliminary Examining Authority	<input type="checkbox"/> other:

The International Bureau of WIPO 34, chemin des Colombettes 1211 Geneva 20, Switzerland Facsimile No.: (41-22) 740.14.35	Authorized officer François BAECHLER Telephone No.: (41-22) 338.83.38
---	---

(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization
International Bureau



(43) International Publication Date
1 February 2001 (01.02.2001)

PCT

(10) International Publication Number
WO 01/08039 A2

(51) International Patent Classification⁷: G06F 17/00 (81) Designated States (national): AE, AG, AL, AM, AT, AU, AZ, BA, BB, BG, BR, BY, BZ, CA, CH, CN, CR, CU, CZ, DE, DK, DM, DZ, EE, ES, FI, GB, GD, GE, GH, GM, HR, HU, ID, IL, IN, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MA, MD, MG, MK, MN, MW, MX, MZ, NO, NZ, PL, PT, RO, RU, SD, SE, SG, SI, SK, SL, TJ, TM, TR, TT, TZ, UA, UG, US, UZ, VN, YU, ZA, ZW.

(21) International Application Number: PCT/US00/20401 (84) Designated States (regional): ARIPO patent (GH, GM, KE, LS, MW, MZ, SD, SL, SZ, TZ, UG, ZW), Eurasian patent (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European patent (AT, BE, CH, CY, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE), OAPI patent (BF, BJ, CF, CG, CI, CM, GA, GN, GW, ML, MR, NE, SN, TD, TG).

(22) International Filing Date: 27 July 2000 (27.07.2000)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data: 09/361,122 27 July 1999 (27.07.1999) US

(71) Applicant (for all designated States except US): ZENECA LIMITED [GB/GB]; 15 Stanhope Gate, London W1Y 6LN (GB).

(72) Inventor; and

(75) Inventor/Applicant (for US only): LERMAN, Charles, L. [US/US]; 501 Bishop Hollow Road, Newton Square, PA 19073-3138 (US).

(74) Agents: BIRD, Donald, J. et al.; Pillsbury Madison & Sutro, LLP, 1100 New York Avenue, N.W., Washington, DC 20005 (US).

Published:

— Without international search report and to be republished upon receipt of that report.

For two-letter codes and other abbreviations, refer to the "Guidance Notes on Codes and Abbreviations" appearing at the beginning of each regular issue of the PCT Gazette.

A2

WO 01/08039

(54) Title: ANALYSIS AND PATTERN RECOGNITION IN LARGE, MULTIDIMENSIONAL DATA SETS USING LOW-RESOLUTION DATA GROUPING

(57) Abstract: Methods, systems and devices for operating on data provide at least one user-defined grouping rule for grouping the data into a user-definable number of groups; and apply at least one of the grouping rules to the data. The data may be in a table, wherein the at least one grouping rule applies to at least one user-selectable column of the table. The grouping rule defines breakpoints corresponding to the user-definable number of groups, and application of the at least one rule to the data divides the data into groups based on the breakpoints. The grouped data is presented in a manner that visually distinguishes the groups, sometimes by coloring an aspect of the data according to the rules.

ANALYSIS AND PATTERN RECOGNITION IN LARGE, MULTIDIMENSIONAL DATA SETS USING LOW-RESOLUTION DATA GROUPING

5 A portion of the disclosure of this patent document contains material which is subject to copyright protection. The copyright owner has no objection to the facsimile reproduction by anyone of the patent document or the patent disclosure, as it appears in the Patent and Trademark Office patent file or records, but otherwise reserves all copyright rights whatsoever.

10 **1. Field of the Invention**

This invention relates to analysis and pattern recognition of data. More particularly, this invention relates to methods, systems and devices and combinations thereof for analysis and pattern recognition in large sets of multidimensional data using low-resolution data grouping.

15 **2. Background**

With the advent of computerization and the low cost of data storage and acquisition, people in many endeavors are now accumulating very large sets of data. For example, scientists in drug and chemical companies now use automation to perform so-called high-throughput screening ("HTS") of chemical compounds. 20 HTS uses automated, relatively low-cost techniques to obtain various items of information about chemical compounds. The goal of using HTS is to obtain information about a very large number of compounds in a quick and relatively low-cost manner. Having accumulated a very large HTS data set, it is necessary to evaluate the data in order to determine which, if any, of the analyzed

compounds warrants further investigation. However, the results of such HTS tend to be very large sets of multidimensional data, on the order of thousands of rows and dozens of columns, and so it is very difficult to make decisions just by looking at the data.

5 In addition to the very large amounts of data produced by HTS, difficulties in existing data handling and analysis methods include the following:

- Data comes from very diverse sources, including HTS laboratories, physical measurements, bio-scientists' laboratories, various computational software programs, etc., and the different sources tend to have very diverse kinds of output including numbers, text, mixed data types, error notations, blank data, replicate data (more than one value per compound), etc.
- Not all sources produce data on the same list of compounds, or in the same order.
- Some data values are misleadingly too precise, i.e., have high relative experimental errors or noise, and can easily be over-interpreted.
- Medicinal chemists have to weigh very different kinds of factors (for example, molecular weight vs. dose-responsiveness vs. ClogP vs. secondary biology vs. selectivity across screens) in trying to determine which are the best compounds or clusters of compounds to which to devote further work.

SUMMARY OF THE INVENTION

This invention solves the above and other problems by providing automated tools to help with and speed up these data handling and analysis processes. These tools embody some assumptions about how the data should be treated by internalizing the most generally acceptable assumptions, but leaving more idiosyncratic decisions to individual users.

A central concept on which this invention is based is grouping data into a relatively small number of categories using low-resolution data grouping. The grouping is visualized by assigning colors to data groups, e.g., in spreadsheets.

Grouping of data potentially changes the precision of the data.

This categorization of data has several major benefits, including:

- creating a visual means of finding data patterns;
- beneficially blurring small variations in numerical data that are, in practice, excessively fine distinctions, possibly due to experimental error; and
- providing, in the colors themselves, a means or “common currency” to evaluate candidates across a wide range of data types.

Accordingly, in one aspect, this invention provides mechanisms to expedite pattern recognition in large sets of multidimensional data, such as those that chemists assemble when evaluating hits from high-throughput screening (HTS) and deciding which ones will get priority for further investigation. In controlled trials, this invention has reduced the time to evaluate real data sets, from days of intense human effort, which is vulnerable to errors due to volume or fatigue, to a few minutes of automation with graphical presentation of results.

It quickly becomes obvious upon using the system that the tools also have value in data-handling areas other than HTS. Examples include selection and management of any kind of tabulated data, e.g., portfolio management for any kind of rated portfolios, selection of drug candidate compounds, selection and management of proteins that are candidates for targets for drugs, selection and management of research projects competing for resources, and evaluating employee performance or job candidates.

The system of this invention includes a new special command menu, a set of graphical user interface worksheets, and action buttons to facilitate the coloring and color analysis processes for the user. While the central process is the data grouping and coloring, there are also new tools for the upstream, or pre-grouping and coloring processes of importing, assembling, regularizing, and characterizing data in a spreadsheet, and for the downstream processes of visualizing, scoring, comparing, and sorting large amounts of color-coded data. The data-grouping and spreadsheet-coloring tool is presently implemented with a flexible, powerful, and convenient user interface that does not require knowledge of spreadsheet macros or of the Visual Basic language (used for the system's implementation).

Accordingly, this invention provides methods, systems and devices for operating on data.

In one aspect, the method of this invention provides at least one user-defined grouping rule for grouping the data into a user-definable number of groups. At least one of the grouping rules is applied to the data. The data may be provided in a table and the grouping rule applies to at least one user-selectable column of the table. In some embodiments, the grouping rule defines breakpoints corresponding to the user-definable number of groups. Application of the rule the

data divides the data into groups based on the breakpoints. The method may include presenting the grouped data in a manner that visually distinguishes the groups. In some embodiments, the grouping rules associate colors with groups and the grouped data is presented with an aspect of the data colored according to the rules.

5

Sometimes the data are in labeled columns in a spreadsheet, and the grouping rule specifies at least one breakpoint and a corresponding color for each range defined by the breakpoint. The grouped data are presented by coloring each data item in one labeled column of the data based on the breakpoint and the corresponding color of the breakpoint.

10

The breakpoints may be numeric or textual values. In some embodiments, the breakpoint is determined automatically based on the data.

Sometimes the data are provided in a table, and backgrounds of table cells are colored according to the rules.

15

The number of groups may be fewer than a number of possible data values.

20

In another aspect, this invention is a method of operating on data by providing at least one user-defined grouping rule for grouping the data into a user-definable number of groups. The grouping rule is applied to the data to generate grouped data. At least one user-defined scoring rule is used to score grouped data according to user-defined scores. The scoring rule is applied to the grouped data to score the grouped data.

In yet another aspect, this invention is a method of operating on data, in which data are grouped by applying to the data at least one user-defined grouping rule for grouping the data into a user-definable number of groups. The grouped

data are scored by applying to the grouped data at least one user-defined scoring rule for scoring the grouped data according to user-defined scores.

In some embodiments the data can be a number of parameters for each of a number of cases and the scoring rule comprises a scoring function of user-selectable parameters and user-defined weights for the selected parameters to be used in scoring the cases. The scoring applies the function to the data to obtain a score for each case. Sometimes the method includes sorting the scored cases by score, individually or by cluster, as described below.

The notion of clustering is that subsets of the various cases may be associated into clusters by having identical entries in any user-selected column of data, known as a clustering column. In some embodiments of the invention, the integrated clusters are treated by averaging the properties of all the cases which comprise each cluster.

Thus, according to aspects of this invention, in order to facilitate analysis and pattern recognition in large, multidimensional data sets, the precision of the data is potentially changed (implemented, e.g., by grouping the data) and then the data are presented for visualization (implemented, e.g., by coloring the data).

BRIEF DESCRIPTION OF THE DRAWINGS

This file contains at least one drawing executed in color. Copies of this patent with color drawings will be provided by the United States Patent and Trademark Office upon request and payment of the necessary fee.

The above and other objects and advantages of the invention will be apparent upon consideration of the following detailed description, taken in

conjunction with the accompanying drawings, in which the reference characters refer to like parts throughout and in which:

FIGURE 1 shows a typical computer system on which the present invention operates;

5 **FIGURE 2** shows an overview of the functionality of the present invention;

FIGURES 3A-3B depict a display of data in a spreadsheet;

FIGURES 4A-4B show a color control rules worksheet according to one embodiment of the present invention;

FIGURES 5A-5B show data coloring rules;

10 **FIGURES 6A-6C** show a data coloring control panel and a flow chart of the data coloring process, respectively;

FIGURES 7A-8B show the worksheet of **FIGURE 3A** and **3B** after various coloring rules in **FIGURE 4A** have been applied;

FIGURES 9A, 9B, 10A, and 10B depict displays of data in spreadsheets;

15 **FIGURES 11A** and **11B** show the form of the cluster control worksheet according to one embodiment of the present invention;

FIGURES 11C-11D shows control panels from the cluster control worksheet of **FIGURES 11A-11B**;

20 **FIGURE 12** shows the enlarging of the cluster starts mechanism according to one embodiment of the present invention;

FIGURES 13A-13D show the application of vertical display re-scaling according to one embodiment of the present invention;

FIGURES 14A-14D and **15A-15B** show the application of the scoring and sorting of clusters according to one embodiment of the present invention;

FIGURES 16A-16N, 16P and 16Q show aspects of the application of the dose-response scoring and estimation of potencies according to one embodiment of the present invention;

FIGURES 17A-17B show the application of the sheet statistics tool according to one embodiment of the present invention;

FIGURES 18A-18D show the application of the scoring and sorting of clusters for the purpose of project prioritization and management according to one embodiment of the present invention;

FIGURES 19-24 show examples of the application of this invention to various types of data; and

FIGURES 25 and 26 show application of an aspect of this invention.

DETAILED DESCRIPTION OF THE PRESENTLY PREFERRED EXEMPLARY EMBODIMENTS

Overview

FIGURE 1 shows a typical computer system 100 on which the present invention operates. The computer system 100 includes a processor (CPU) 102 connected to a memory system 104 and a display 106. The computer system also includes various input devices including a keyboard 108 and a mouse 110 or other pointing device. Internal storage 112 (e.g., a hard disk, a CD ROM and the like) and external storage 114 (such as a floppy disk, CD ROM and the like) are also provided.

Various aspects of this invention are implemented as computer software programs or algorithms 116 which run on the computer system 100. The software programs 116 can reside in the internal storage 112, the external storage 114,

and/or in the memory 104. The software programs 116 operate on data 118 which is provided, e.g., on the external storage 114. The software programs 116 operate in a standard and known manner by being executed on the processor 102 of the computer system 100.

5 In some embodiments of the present invention, the user can create and modify various executable rules 120 which can operate on the data 118. For the sake only of explanation, the rules 120 are depicted separately from the data in the figures. However, as explained in more detail below, some or all of the rules 120 can be part of the data 118.

10 In preferred embodiments, the computer system 100 is capable of running the spreadsheet program EXCEL™ 95 (hereinafter "EXCEL") from Microsoft Corporation, and the software computer programs 116 are written in Microsoft Corporation's Visual Basic (hereinafter "VB") and are provided as an add-in to EXCEL. A single copy of software thus serves all data files on a particular machine. To conserve EXCEL resources, in some embodiments, the package self-15 installs the add-in when the user opens a data file, and un-installs the add-in when the last data file in memory is closed.

20 In a preferred embodiment, this invention works entirely within the environment of EXCEL. EXCEL structures data files as workbook files which contain worksheets. The programs 116 of this invention consist of special EXCEL worksheets, called control sheets, on which input data is written by the user into designated labeled cells. The control sheets are part of the same EXCEL workbook file as the data. The control sheets also contain action buttons to execute the various procedures associated with this invention. The rules 120 are formed by 25 setting various parameters in the control sheets.

When the workbook file is saved, the parameters (for the rules 120) are stored on the control sheets along with the data, and they can be modified and/or re-executed at any time without having to re-enter anything. The results of operations are automatically written as worksheets in the same workbook file, 5 providing a convenient, integrated data environment in a single file.

The system according to the present invention operates, in one aspect, in accordance with **FIGURE 2**. Recall that the user's aim is to perform analysis and pattern recognition in large, multidimensional data sets using (potentially low resolution) data grouping. To this end, the user and/or the system will create rules 10 for coloring and presenting the data. First (at 122) a user creates and organizes the data 118. Various tools (discussed below) are provided to aid in the creation and organization of the data. Then (at 124) the user creates rules 120 for operating on the data 118. The rules 120 can be created before or after the data 118, rules can be reused for different sets of data and multiple rules can apply to the same data. 15 The creation and operation of rules are discussed in greater detail below. Once the data 118 and the rules 120 are created, the user then selects some (or all) of the rules to apply to the data (at 126). Specifically, the user groups and thereby colors the data according to selected rules. With the data grouped and colored according to the rules, the user can then perform group/color-mediated data mining (at 128).

20 **FIGURES 3A-3B** show views of the program of this invention in operation with a sample EXCEL sheet 300, denoted "DEMO 1" (302) containing data (not all the data in the sheet is visible). The views of EXCEL worksheets shown in the various figures and examples that follow are the views that are presented on the display 106 of the computer system 100. Sheets in an EXCEL workbook are 25 labeled with tabs at the bottom of the worksheet. The data on the "DEMO1" sheet

300 consists of eight columns of data for each of a number of compounds. The compounds are denoted "Cmpdxx", where "xx" ranges from "01" to the number of compounds. In **FIGURE 3B**, the last compound visible on the data sheet is "Cmpd58". The eight columns are headed:

- 5 1. "Cmpd" (column A);
2. "Series" (column B);
3. "Test1" (column C);
4. "Test2" (column D);
5. "Test3" (column E);
- 10 6. "HTS SPA Dose-Resp % Inhib @3x10-6M" (column F);
7. "HTS SPA Dose-Resp % Inhib @ 1x10-6M" (column G);
8. "HTS SPA Dose-Resp % Inhib @ 3x10-7M" (column H); and
9. "HTS SPA Dose-Resp % Inhib @ 1x10-7M" (column I).

In addition to the "DEMO 1" worksheet 300, the EXCEL workbook shown in **FIGURES 3A and 3B** has seven other worksheets, denoted "DEMO 2" 304; "DEMO 3" 306; "clusterinfo DEMO" 308; "Append Control" 310; "Color Control" 312 and "Cluster Control" 314. The last three worksheets, denoted respectively "Append Control"; "Color Control" and "Cluster Control," contain various rules and controls (to be discussed below). The data in worksheets denoted "DEMO 1" 302; "DEMO 2" 304; "DEMO 3" 306; and "clusterinfo DEMO" 308 correspond to data 118 (**FIGURE 1**) and the controls or rules in the worksheets denoted "Append Control" 310; "Color Control" 312, and "Cluster Control" correspond to the rules 120 (**FIGURE 1**).

25 **FIGURES 4A-4B** show a color control rules worksheet (312, denoted "color control") according to the present invention, as displayed on display 106 of the

computer system 100. The color control worksheet 312 is shown with some rules already in place, i.e., having values set, and other rules left blank. These rules are shown as examples only, and, as with any of the other types of rules, any or all of the rules can be set by the user. A typical data coloring rule 130 is shown in **FIGURE 5A**. The rule 130 has already been set up and operates on the appropriate data when selected by a user (using mouse 110, **FIGURE 1** or some other pointing device) in the area 132 marked "Click here to run these". The rule 130 (as with all of the preferred color control rules) has four parts, namely the name of the sheet 134 containing the data on which the rule is to operate ("DEMO 1" in the example of **FIGURE 5A**); the columns 136 of data of the sheet on which the rule is to operate ("E" in the example of **FIGURE 5A**); the number of colors 138 to be used by the rule 130; a number of breakpoints 140 (denoted "break 1" to "break 4" in the example of **FIGURE 5A**); and a corresponding number of colors 142 for each range defined by the breakpoints (denoted "color 1" to color 4" in the example of **FIGURE 5A**). Actually, as explained below, the number of breakpoints is one less than the number of colors.. In the specific example shown in **FIGURE 5A**, the rule has three breakpoints of 1, 5 and 10, defining four ranges with four corresponding colors 142, namely light green, yellow, orange and red. Preferably the named colors are also depicted in the actual colors, so that, in this example, the background of the word "lightgreen" is shown in light green, the background of the word "yellow" is shown in yellow and so on.

In this invention it is preferable to show data and meta-data (headings etc.) in color. In some embodiments, the coloring is implemented by showing a background area of the text representing the data in the appropriate color. Sometimes the actual text representing the data is shown in the appropriate color.

In presently preferred embodiments, the font color is only changed in cases where necessary to improve contrast with the background color for readability. Only two font colors, dark (black) and light (pale gray), are used in the presently preferred embodiment. Combinations of both approaches can be used. For example, the background section of the word "yellow" is preferably shown in the color yellow. It is also possible to show the word itself, i.e., the font, in the color yellow, as long as that color is distinguishable from the background.

5

The particular rule 130 shown in FIGURES 4A and 5A, operates as follows,

when selected:

10

In sheet "DEMO 1" 302, in column E, values less than or equal to 1 (break 1) are colored light green (color 1); values in the range 1 to 5 (between break 1 and break 2) are colored yellow (color 2); values in the range 5 to 10 (break 2 to break 3) are colored orange (color 3); and values greater than 10 (break 3) are colored red (color 4).

15

Another typical data coloring rule 130-1 from the color control sheet 312 is shown in FIGURE 5B. The rule 130-1 is set up to operate on columns "C" and "D" of sheet "DEMO 1". The rule 130-1 uses three (3) breakpoints (break1=0.1, break2=1 and break3=5) defining four ranges with four (4) corresponding colors ("lightgreen", "yellow", "orange", and "red").

20

The results of applying the rule 130-1 of FIGURE 5B to the data in sheet "DEMO 1" (302, FIGURE 3) are shown in FIGURES 7A-7B. As can be seen from FIGURES 7A-7B, after application of the rule 130-1, all of the data in columns C and D of the sheet labeled "DEMO 1" has been colored according to the rule. Specifically, data having a value less than or equal to break 1 (0.1) have been colored light green; data values in the range between break 1 and break 2 (0.1 to

25

1) have been colored yellow; data values in the range between break 2 and break 3 (1 to 5) have been colored orange; and data values greater than break 3 (5) have been colored red.

The results of applying all of the other color control rules shown in
5 **FIGURES 4A-4B** to the data in sheet “DEMO 1” are shown in **FIGURES 8A-8B**.
The rules can be applied individually (as shown above with respect to
FIGURES 7A-7B), or they can be all be applied at the same time. In order to apply
all rules to a particular data set (sheet), each rule can be individually selected or
the area labeled “RE-RUN ALL RULES FOR SHEET NAMED DEMO 1” (on the right
10 side of **FIGURE 4A**) can be selected. Note that if two rules apply to the same
column of the same sheet, the second rule run on that column will override the
first rule run on that column.

To create a coloring rule a user performs the following (with reference to
15 **FIGURE 6B**):

(1) Select the “COLOR CONTROL” sheet **312** and pick a control panel on
that sheet to use (an empty panel or one containing a rule no longer
needed) (at **600**). All control panels on a sheet can be cleared by
clicking the button labeled “CLEAR ALL ENTRIES ON THIS SHEET”
20 (318 in **FIGURE 4A**).

(2) In the selected control panel, enter the name of the sheet to be
colored (at **602**).

(3) In the selected control panel, enter a column or columns (at **604**).
For multiple columns, either list them separated by commas, or use
a colon or hyphen to denote ranges, or some combination. For
example, “A:D,F” means columns A,B,C,D, and F. To aid in
25

WO 01/08039

choosing columns, the user can right-click on the cell containing the name of the data sheet, and pick “Open Twin Screen” from the shortcut menu that appears, to create a special dual display. This also creates a “Close Twin Screen” button to go back.

5 (4) Choose a number of colors to use (at 606), either by entering the number of colors or by repeatedly clicking the gray button adjacent to the cell labeled “# of colors”. In preferred embodiments, the system allows for five breakpoints and six colors per rule. Accordingly, the numbers will cycle from 1 to 6, and various cells below them will be blacked out accordingly.

10 (5) Enter the breakpoints that define the color groups (at 608), in any of three modes:

15

a) Numeric data, manual mode: enter numbers to form the breakpoints, i.e., the boundaries between the color groups, one less than the number of colors, in increasing numerical order. Cells whose values exactly equal a breakpoint value will be colored with the lower group (breakpoint 1 is colored with color 1, etc.)

20

b) Numeric data, automatic mode: enter either “value”, “log”, or “count” as the first breakpoint. If multiple columns have been chosen, the user must also enter “yes” or “no” opposite “Re-scale all?” at the bottom of the panel, to indicate whether each column should get its own auto-breakpoints, or whether the

25

auto-breakpoints of the first column (first in list in the rule, not first on the data sheet) should be used for all.

This mode reports information about the 5 breakpoints it determines, and thus could also be used to explore the distribution of numerical values in a column prior to a final *manual* breakpoint selection.

c) Text data: enter the strings to be matched and 10 colored, in preferred embodiments, up to five (5) in number. Matching is case-insensitive unless the string is enclosed in double quotes (" and "); otherwise, no quotation marks are necessary.

Several special text strings act as operators if entered as 15 the first word in a rule cell:

<u>Rule Entry</u>	<u>Meaning</u>
(OPERATORS need not be uppercase—here only for emphasis)	
test string	color data cell if its whole content matches the test string
NOT test string	<ul style="list-style-type: none"> color data cell if its whole content does not match the test string; will not color numeric cells
CONTAINS test string	color data cell if contains the test string as a substring anywhere
NOTCONTAINS test string	<ul style="list-style-type: none"> color data cell if it does not contain the test string anywhere; will not color numeric cells

<u>Rule Entry</u>	<u>Meaning</u>
(OPERATORS need not be uppercase—here only for emphasis)	
BEGINS test string	color data cell if it begins with the test string
ENDS test string	color data cell if it ends with the test string
* (an asterisk)	(wildcard) color data cell containing any data, including numeric cells
BLANK	color data cells that are blank

Using quotes to force matching to be case-sensitive also works with strings that follow an operator.

It is possible to construct a text-coloring rule in

5

which certain cells may satisfy more than one of the

“breakpoint” values. For example, if a rule says that

“active” is colored green and “contains act” is

colored red, then the word “active” in a cell would

satisfy both. In such cases, the system colors the

cell according to the *first* condition satisfied on the

list of breakpoints. This dependence on the

order can be used advantageously to achieve

complex coloring conditions. The sequence of

conditions can be considered as a series of filters,

through which only the as-yet-uncolored cells fall

10

through to the next decision.

15

(6) Enter the names of the colors to use (at 610), in the order corresponding to the breakpoints. A display of color samples is provided at the right side of the Color Control sheet 312. A user need only enter the name, and the appropriate cell will become colored when the tool is executed. If the user wants the color to display immediately, he can copy and paste the sample cell into the rule's color cell. A special pseudo-color named "SKIP" is used to tell the system not to color the cells whose data falls in this group.

5

(7) When the rule has been created, the user executes the rule by

10

selecting the rule's "CLICK HERE TO RUN THESE" button on the panel filled in (at 612, FIGURE 6C).

15

(8) To create different coloring rules for other columns, repeat the

above in additional control panels. If the user runs out of control panels, he can create more control panels by copying an existing one and pasting it onto a blank section of the color control sheet.

To the extent that a single panel cannot hold all the requirements for a particular rule, a user can combine two or more panels to create a particular rule. For example, if a user needs ten (10) breakpoints, two panels can be used.

20

With reference to the coloring rule is shown in FIGURE 6A, once the rule has been set, a number of parameters are stored in the system. The parameters are "sheet name" ("DEMO 3" in FIGURE 6A), column specification, number of colors, array of breakpoints, array of colors, and multicolumn scaling mode.

The data coloring mechanism operates as follows, with reference to the flowchart of **FIGURE 6C.:**

1. The user enters the parameters into a rule panel on a "COLOR CONTROL" worksheet 312, e.g., as described above with reference to the panel of

5 **FIGURE 6A.**

2. The user selects (clicks) the activation button (labeled "Click here to run these") on that panel (at 612). This causes the system to:

(A) Read and interpret the parameters from the panel (at 614).

10 The system can identify which button was clicked using the Visual Basic ("VB") "caller" property, described in more detail below. The parameters are then read based on the identity of the cell location of the button using the VB "TopLeftCell" property. The system retrieves the parameters (sheet name, column specification, number of colors, array of breakpoints, array of colors, and multicolumn scaling mode) from cells in this panel by relative reference to the button cell.

15 (B) Next, determine the mode of the coloring rule (at 616) (i.e., numeric v. text or manual v. automatic, and, if automatic, which of value, log or count). This uses the analysis of the first breakpoint entry.

(C) Compile a list of the columns specified in the "column specification" parameter (at 618). This is done by scanning the various areas contained in the selection, as follows:

```
20 For Each singlearea In Selection.Areas
    For Each c In singlearea.Columns
        If Not CountEmpty Then
            lr = LastRowInColumn(c.Column)
        End If
```

```
If Not CountEmpty And lr = 0 Then
    'skip this empty column
    ncols = ncols - 1
Else
    ' add this column to the list
5
    icol = icol + 1
    colnumarray(icol) = c.Column
End If
Next c
Next singlearea
```

10

(D) If an auto-breakpoint mode is being used (determined at 620), analyze the data values to determine the breakpoints (at 622). This is done by:

15

- (i) Collecting statistics on the data distribution in each specified column; and
- (ii) Calculating the automatic break points for the appropriate mode. For example, the auto-value breakpoints are determined as follows:

```
20    If breakmode = "VALUE" Then
        interval = (maxvalue - minvalue) / ncolors
        break(0) = minvalue
        For ibreak = 1 To ncolors - 1
            break(ibreak) = break(ibreak - 1) + interval
25    Next ibreak
```

- (iii) Displaying the results for user approval or cancellation.

(C) Loop through the cells in the chosen columns on the chosen 30 worksheet (at 624).

(D) Compare each cell's value to the list of breakpoints (at 626). If the coloring rule is in text mode, use the meanings of the special breakpoint operators ("contains", "blank", asterisk "*"; or quotation marks).

5

(E) When a match is found, apply the appropriate color (at 628).

The code below illustrates the processes (D) and (E) for numeric

breakpoints:

10

```
For Each cell In Range(Cells(StartColoringRow, colnum) ,  
                      Cells(FinishColoringRow, colnum) )  
    cvalue = cell.Value  
    colored = False  
    15    If IsNumeric(cell.Value) Then  
        If Not IsEmpty(cell) Then  
            ' (have to test both because EMPTY is numeric)  
            For ibreak = 1 To ncolors - 1  
                If cvalue <= break(ibreak) Then  
                    If Color(ibreak) <> "SKIP"  
                        20                    cell.Interior.ColorIndex =  
                        Color(ibreak)  
                        Call TextContrast(cell)  
                    End If  
                    colored = True  
                Exit For  
            End If  
        Next ibreak  
    If colored = False Then  
        ' Not hit yet? Must be top category, so:  
        30        If Color(ncolors) <> "SKIP"  
            cell.Interior.ColorIndex = Color(ncolors)  
            Call TextContrast(cell)
```

```
        End If
        colored = True
    End If
    End If
5     Else      ' not numeric - just don't color it
End If
Next cell
```

The operation of the data coloring tool of this invention will now be described in greater detail. Each coloring rule is provided in a coloring control panel that has the general form of a coloring rule as shown in **FIGURE 6A**. In one preferred embodiment, each coloring control panel **144** is a double-outlined unit, sixteen (16) cells high by two (2) cells wide. As noted above, a user is provided with coloring control panels on the color control worksheet **312**. A user can use the coloring control panel **144** to set the sheet and column(s) on which the rule is to operate, the number of colors, the various break points and the colors associated with those breakpoints. The sheet is set by entering its name into the cell **146** adjacent the cell labeled "sheet". The column (or columns) on which the rule is to operate is (are) set by entering its (their) name in the cell **148** adjacent the cell **15** labeled "column(s)". The number of colors is set by the user by selecting the cell **150** adjacent to the cell labeled "# of colors". Each time the cell **150** is selected it increases the number of colors, up to a maximum of six (6), i.e., rotating through the values 1 to 6. I.e., when the cell **150** shows a "6" and is selected, it reverts back to "1". That is, selecting the cell **150** causes the value in the cell to cycle **20** from "1" to "6" and then back to "1".

25 A Visual Basic ("VB") macro function ("CallColorColumn") is associated with the top cell **152** of the control panel **144**. When the cell **152** is selected by

the user (with the mouse 110 or the like), the function associated with that cell is executed by the computer (CPU 102). In the presently preferred embodiments, the *CallColorColumn* function extracts the button name of the cell 152 and then calls a second function ("CallColorColumnSubroutine") with that button name as one of the parameters. The function *CallColorColumnSubroutine* takes three parameters, namely *ButtonName*, *StartColoringRow*, and *FinishColoringRow*.
5 The two parameters *StartColoringRow*, and *FinishColoringRow* are optional.

First, the function *CallColorColumnSubroutine* determines what specific values to use for the coloring by reading them from the control panel 144. Since 10 the values are all in fixed positions relative to the selected button cell 152 that initiated the call to the function *CallColorColumn*, the values can be determined once the location of that button cell 152 has been determined. This is done using the following Visual Basic code:

```
15 Sheets("Color Control").Activate  
headingrow =  
    ActiveSheet.Buttons(ButtonName).TopLeftCell.Row  
headingcol =  
  
    ActiveSheet.Buttons(ButtonName).TopLeftCell.Column
```

20

Note that if the function *CallColorColumnSubroutine* was called from another sheet (not "Color Control") then this method will not find it.

The various parameter values are then read as follows:

Sheet name:

```
25 datacol = headingcol + 1  
sheetname = Trim(Cells(headingrow + 1, datacol).Value)
```

If there is no sheet named “sheetname” an error function is called.

Generally, in preferred embodiments, a great deal of error checking takes place to ensure that the user is provided with a friendly and useable interface to the program. Most of the error checking is not mentioned in this description, however, one skilled in the art would know what kinds of error checking to implement in order to provide a user-friendly working environment.

5

The column(s) to be colored are specified by:

```
Cells(headingrow + 2, datacol).Value
```

10

The number of colors is specified by the variable *ncolors*, where:

```
ncolors = Cells(headingrow + 3, datacol).Value
```

15

Within the function *CallColorColumnSubroutine* there are two arrays, named *break* and *color*, which are used to store the breakpoints and colors, respectively. The first breakpoint is set as follows:

```
break(1) = Cells(headingrow + 4, datacol).Value
```

20

The value of the first breakpoint is used to determine the break mode (“NUMERIC”, “VALUE”, “LOG”, or “COUNT”). If *break(1)* (as determined above) is numeric, then the mode is set to “NUMERIC”, otherwise, if *break(1)* is one of “VALUE”, “LOG”, or “COUNT”, then the break mode is set to that mode, otherwise the break mode is set to “TEXT”.

25

Next, the function determines whether multiple columns were specified, in which case it determines whether the user selected to re-scale all the columns.

The user's re-scale selection is determined by:

```
5      rescale_all_string = Cells(headingrow +
      15, datacol).Value
```

Now the rest of the breakpoints (if any) are read. If the break-mode is "AUTO" then the breakpoints are set as follows:

```
10     For i = 2 To lastbreaknum
          break(i) = Cells(headingrow + 3 + i, datacol).Value
```

15 Various possible errors are checked for. E.g., if any breakpoints are missing (i.e., if *break(I)* is empty, the user is notified. Also, if the break mode is "NUMERIC" and non-numeric breakpoints are set, the user is notified. If numeric breakpoints are not in increasing order, the user is notified. As noted above, generally, in preferred embodiments of the present invention, a great deal of error checking is performed on all user inputs to ensure that the values are correct and consistent. Most error checking is omitted from this description.

20 The *CallColorColumnSubroutine* maintains an array, *colortname*, which maps integers to colors. In preferred embodiments, there are fifty six (56) colors available. To use the higher numbered colors, the computer's video card must be set appropriately. Using the *colortname* array, the program next associates the user provided color names with index numbers. Specifically, for each of the (up to six in a preferred embodiment) colors specified, the user specifies an actual color name, denoted *cname*. This name is determined for each color by:

```
25     For j = 1 To ncolors
          cname = Cells(headingrow + 8 + j, datacol).Value
```

WO 01/08039

The interior of each color-specifying cell is then colored by the appropriate (selected) color by setting the color property (*Interior.ColorIndex*) of the cell:

```
Cells(headingrow + 8 + j, datacol).Interior.ColorIndex =
    5
    Color(j),
```

where the value of the variable *j* ranges from 1 to *ncolors*.

Then the cell is further processed by a function *TextContrast*.

```
Call TextContrast(Cells(headingrow + 8 + j, datacol))
```

With the parameters read and checked, the system is ready to process and

10

color the selected sheet (specified at cell 146 in **FIGURE 6A**). The selected columns (specified in cell 148 in **FIGURE 6A**) in the selected sheet are processed one-by-one by the following program code:

```
Call ParseInput(InString, inspecifier)
For Each singlearea In Range(inspecifier).Areas
    15
    For Each c In singlearea.Columns
        colnum = c.Column
        Call ProcessOneColumn(colnum, ncolors, break,
            Color, breakmode,
            rescale_all, sheetname,
            StartColoringRow,
            FinishColoringRow)

        Next c
    Next singlearea
```

20

The processing performed by the function *ProcessOneColumn* is as follows: The column designated by *colnum* on sheet *sheetname* is to be colored according to the breakpoints in the array *break* and the colors in the array *colors*. The designated column is colored from the row corresponding to *StartColoringRow* to the row corresponding to *FinishColoringRow*. Note that the

30

WO 01/08039

function *ProcessOneColumn* is also provided with the break mode and the variable *rescale_all*.

Function *ProcessOneColumn* first calculates the automatic breakpoints, if necessary. Note that automatic breakpoints are determined from the whole column, even if this call says to color only a limited range of rows. If the value of *breakmode* is "VALUE" or "LOG" and the value of *rescale_all* is set to "True" Or the value of the first breakpoint (*break(1)*) is set to "VALUE" or "LOG", the program calls the function *AutoValueBreakpoints* as follows:

```
Call AutoValueBreakpoints (colnum, colletter, ncolors,
                           break, Color, breakmode, rescale_all).
```

10

Otherwise, if the *breakmode* is set to "COUNT" and the value of *rescale_all* is set to "True" or the first breakpoint (*break(1)*) is set to "COUNT", then the program calls the function *AutoCountBreakpoints*, as follows:

```
Call AutoCountBreakpoints (colnum, colletter, ncolors,
                           break, Color, breakmode, rescale_all, sheetname).
```

15

With the breakpoints calculated, the columns are colored according to the type of breakpoints specified by the user. Specifically, when the *breakmode* is any one of "VALUE", "COUNT", "LOG", or "NUMERIC", the system executes a function *ColorNumericColumn*. On the other hand, when the *breakmode* is "TEXT", the system executes a function *ColorTextColumn*. The VB code for this is as follows:

```
25      Select Case breakmode
              Case "VALUE", "COUNT", "LOG", "NUMERIC"
```

```
Call ColorNumericColumn(colletter, ncolors,
break, Color, StartColoringRow, FinishColoringRow)
Case "TEXT"
    Call ColorTextColumn(colletter, ncolors, break,
5 Color, StartColoringRow, FinishColoringRow)
End Select
```

Then, when the rule in control panel 144 is selected for execution, the rule is applied to the selected column(s) (denoted in cell 148) of the named sheet (in cell 146). For each column in the named sheet, the value in each cell is compared to the various breakpoints and the cell is colored corresponding to the appropriate 10 breakpoint.

Examples of the application of various coloring rules in the "COLOR CONTROL" worksheet of **FIGURES 4A-4B**, are shown with reference to the data in 15 worksheet "DEMO 2" (depicted in **FIGURES 9A, 9B, 10A and 10B**).

Color-Mediated Data Mining

As noted above with reference to **FIGURE 2**, once the data have been colored according to the user-selected rules (at 126), the user can then perform 20 color-mediated data mining (at 128). The presently preferred embodiment of this invention provides five mechanisms (each discussed below) for color-mediated data mining, namely mechanisms to:

1. enlarge/shrink cluster starts;
2. vertically re-scale the display;
3. score and sort clusters; and
4. score and sort dose-response data.

The following discussion refers to the cluster control worksheet which is shown in **FIGURES 11A-11B**.

1. Enlarge/Shrink Cluster Starts

5 The “Enlarge Cluster Starts” mechanism highlights the first row of each cluster in clustered data by enlarging the font of the cell containing the cluster number or label, thus enabling size reduction of the spreadsheet for the user to focus on the color patterns. When the cell height is dramatically reduced in order to see more cells on a screen or printed page, this enlargement allows the user to 10 still read the label at the beginning of each cluster. The mechanism takes user input from a *Cluster Control* worksheet. A corresponding mechanism (“SHRINK CLUSTER STARTS”) allows for undoing the enlarging. This mechanism handles cluster numbers or textual labels. Any column can be designated as the cluster labels to be processed.

15 Operation of the mechanism is as follows:

20

- (1) From the “CLUSTER CONTROL” sheet **314** pick a control panel to use (one which is empty or one containing inputs no longer needed). On this sheet, a single control panel extends vertically through the black, blue, red, and green sections, and provides input information for several tools.
- (2) In the blue section, enter a sheet name and the column to be considered as the cluster labels.
- (3) Click either the blue-text “Enlarge Cluster Starts” or “Shrink Cluster Starts” button.

The program code accomplishes this by scanning the column of cluster labels, identifying any entries that are different from the one immediately above, and enlarging them. Code that carries out this function is shown below:

```
5  For Each c In Range(Cells(3, colnum), Cells(lastrow,
    colnum))
    irow = c.Row - 1
    icol = c.Column
    If c.Value <> Cells(irow, icol).Value Then
10      c.Font.Size = bigfontsize
      '           Rows(Irow + 1).RowHeight = bigrowheight
      End If
    Next c
```

Example

15 An example of the application of the enlarge cluster mechanism is shown in **FIGURE 12** which shows the application of a rule (shown in the control panel **FIGURE 11C**) from the cluster control worksheet in **FIGURE 11B** to the data of worksheet “DEMO 2” as shown after coloring in **FIGURES 10A-10B**. As shown in **FIGURE 11C**, the rule is to be applied to column B of sheet “DEMO 2”.

20

2. Vertical Display Re-Scaling

25 The vertical re-scaling mechanism operates by taking a user-provided scale factor and then changing height of data rows to facilitate visualization of large-scale color patterns. The mechanism leaves column heading heights and column widths unchanged. This makes headings remain readable and greatly simplifies examining long columns of data for color patterns.

FIGURES 13A-13D show the application of the vertical display re-scale mechanism according to the present invention. **FIGURES 13A-13B** show some of the data in the worksheet labeled “DEMO 3” 306 (**FIGURE 13A** shows the first thirty eight or so elements and **FIGURE 13B** shows the remaining elements of that worksheet). As can be seen from the figures, the worksheet “DEMO 3” 306 has three hundred and twenty eight (328) data entries (in rows 2-329). The user can vertically scale the display by selecting “Re-scale Vertical” from the system’s special menu or by pressing a particular control key sequence (e.g., “CNTL-SHIFT-V” in a preferred embodiment). This presents the user with a dialog box 318, as shown in **FIGURE 13C**, which asks the user to enter a scaling factor relative to the current size. The user enters a scaling factor to enlarge or reduce or restore the display. In the example shown, the user enters a scaling factor of 0.1 which produces the vertically scaled display shown in **FIGURE 13D**.

Vertical scaling allows a user to get an overview of the data, based on the 15 coloring.

The portion of the program code presented below carries out the central function of the vertical display rescaling mechanism:

```
rowspec = "2:" & lastrow ' leaves the headings unchanged,  
20 i.e., readable  
If factor = -1 Then  
    Rows(rowspec).Rows.AutoFit  
Else  
    For irow = 2 To lastrow  
        25 Rows(irow).RowHeight =  
            Rows(irow).RowHeight * factor  
    Next irow  
End If
```

After execution of the rescaling mechanism, as can be seen in

FIGURE 13D, the height of each row (except the heading rows) has been scaled by factor, 0.1 in the example shown. In this manner, all rows of the data are made

5 visible on a single page, thereby facilitating data analysis.

3. Scoring and Sorting Clusters

Scoring and sorting clusters assign numerical scores to the color patterns of individual rows or clusters of rows, thereby enabling comparison and sorting of the clusters by score.

10 The scoring and sorting mechanism accepts user's designations of colors and corresponding relative scores. It handles cluster numbers or textual labels. Any column can be designated as the cluster labels to be processed. The mechanism scores a user-selected list of columns of data, with user-assigned relative weights, which need not be equal for all columns.

15 User input is taken from a *Cluster Control* worksheet 314 (see **FIGURES 11A and 11B**), which stores any number of parameter sets, each one with a user-specified name.

The input data is automatically sorted by cluster label before starting, in order to group the clusters together in case the user has previously sorted the data 20 by some other criterion. Then scores are normalized to remove the effects of cluster size, absolute magnitude of scoring points chosen, and absolute size of weights chosen. The results are written to two new worksheets without altering the original data sheet. The first derived sheet is for the numerical scores; the second is like the original, but has the clusters sorted into descending score order,

so that the "best" are at the top, removing the need to visually scan a long colored worksheet. The derived output sheets have names that indicate their source data sheet and the name of the parameter set used for scoring. At the user's option, the system reversibly hides the un-scored columns in the cluster-sorted output sheet, 5 focusing attention on the data that were used in scoring.

In preferred embodiments, the system detects uncolored cells in the data and offers the user two programmed modes of dealing with them, (uncolored = entry on user's list of scores or uncolored = "average of other colors in row"), or the option of stopping to color them manually.

10 If the user designates a column of individual compound labels as the "cluster labels," then the system compares single compounds rather than clusters.

The mechanism operates as follows, with reference to **FIGURES 11A-11C**.

(1) On the "Cluster Control" sheet **314** the user picks a control panel (e.g., 15 panel **1100**) to use (a panel which is empty or one containing non-needed inputs). On this sheet, a single control panel extends vertically through the black, blue, red, and green sections, and provides input information for several tools.

(2) In the top black section **1102** of the selected control panel **1100**, the user gives this new parameter set a name if not already done. The 20 name will be used to label the outputs.

(3) In the blue section, the user enters a sheet name (in **1104**) and the column (in **1106**) to be considered as the cluster labels. Note: To score each compound separately rather than in clusters, enter a column with individual compound labels as the "Cluster Col."

(4) The red section of the control panel is divided into two parts, with its action button **1108**, with red text “Score and Sort Clusters”, in the middle. Above the button, enter the names of the colors **1110** to be assigned point scores, along with their corresponding point scores **1112**. The scores are arbitrary and relative; they will be normalized by the system as necessary. However, a user should be sure always to assign higher point scores to colors which denote favorable values, and lower point scores to colors which denote unfavorable values. The cells with entries need not be colored, and need not be in score order, because the system will color and sort these cells when run.

When assigning point values, a user should be aware that uncolored cells (which are most likely blank, i.e., unknown data) may have quality values above or below those that contain grouped and colored data.. The user may decide that some of the colored groups are “better” or “worse” than data being unknown, and can assign a score to the color “NONE” accordingly.

(5) Below the “Score and Sort Cluster” button **1108**, the user enters the columns **1114** to use for scoring (using the same syntax as for the Data Coloring) and their corresponding relative weights **1116**. The numbers for weights are arbitrary and relative; they will be scaled by the system as necessary. Note that a line with multiple columns will assign the entered weight to *each* of the columns.

(6) The user selects (clicks) the red-text “Score and Sort Clusters” button **1108**.

(7) When the scoring and sorting tool runs (on the system 100), if the system detects uncolored cells in the data, the user will be offered two modes of dealing with them automatically, or a third manual option of stopping to color them. The two modes are:

5

- "Use score for the color "none" on my list"
(RECOMMENDED)
- "uncolored = average of other colors in row".

10 (8) The program then scans the chosen columns in each row and adds up the chosen column's color scores for that row. These scores are then averaged for each cluster of rows, as defined by the user-selected "cluster column." The VB program code which accomplishes this is as follows:

```
15  For icol = 1 To ncols
      colorcode =
      Cells(irow, colnum(icol)).Interior.ColorIndex
      colorfound = False
      ' Add up the weighted scores
20  For j = 1 To ncolors
      If (icolor(j) = colorcode) Then
          jscore = j
          colorfound = True
          Exit For
25  End If
      Next j
      If colorcode = xlNone And treatblanks = "AVERAGED"
Then colorfound = True
      If Not colorfound Then
          Cells(irow, colnum(icol)).Select
30  If colorcode = xlNone Then
```

```
        thiscname = "none"
    Else
        Call ColorNameToIndex(thiscname, colorcode, True)
    End If
5      addscore = score(jscore) * colweight(icol)
      cmpdscore = cmpdscore + addscore
      ' Next IF-THEN-ELSE block is
      ' special calculations for the "averaged" mode
      If colorcode = xlNone Then
          lostweight = lostweight + colweight(icol)
          minscore = Application.Min(minscore, 0)
          maxscore = Application.Max(maxscore, 0)
      Else
          cmpdscore2 = cmpdscore2 + addscore
15     End If
      Next icol
```

(9) The scores are then normalized for the various cluster sizes (number of rows per cluster), and scaled to a value of one hundred (100) for a row which is colored entirely with the user's highest-scoring color and a value of zero for a row that is colored entirely with any color to which the user has assigned a score of zero.

```
20      If clusterscore(icluster) = 0 Then
            ' do nothing
        ElseIf clusterscore(icluster) > 0 Then
            clusterscore(icluster) =
                100 * clusterscore(icluster) / (nrows * maxscore)
        ElseIf clusterscore(icluster) < 0 Then
            clusterscore(icluster) =
                100 * clusterscore(icluster) /
                (nrows * (-minscore))
30      End If
```

(10) The results are presented as two newly inserted worksheets. The first is named by appending the word "SCORES" to the name of the original data sheet, and contains a list of the clusters with their sizes and scores.

5 (11) The second new sheet is named by appending the word "SORTED" to the name of the original data sheet. The "SORTED" sheet contains a copy of all the original data and coloring, but with the rows re-ordered to place the highest-scoring clusters at the top, and all the clusters in descending score order from there down.

10 (12) The user has two additional options regarding the appearance of the "SORTED" sheet: (a) a column containing the numerical scores can be added; and (b) the columns that were not used in the scoring can be hidden, so that only the ones actually used remain visible.

15 An example of user provided data is shown in the control panel in **FIGURE 11D** which is taken from the cluster control worksheet shown in **FIGURE 11A**. As shown in **FIGURE 11A**, the parameters are stored with the name "Cmpd" **1102**. The scoring a sorting parameters in the control panel **1100** of **FIGURE 11D** give the color red a score of "-1", orange has a score of "0", yellow has a score of "1" and light green has a score of "2". Columns C and D have 20 relative weights of "1", as does column E.

Note on the output of score and sort clusters: The system inserts two new sheets after the data (see, e.g., **FIGURES 14C-14D**). The first added sheet contains two score columns: the scores generated by *both* of the auto modes (uncolored = zero and uncolored = average), but the one not selected will be gray. The scores

are on a scale of “-100” to “+100”, where a score of “-100” means that all cells had the maximally negative score available, and a score of “+100” means that all cells had the maximally positive score available. The second added sheet has clusters sorted according to the *one* auto mode chosen when the tool ran. The 5 routine offers to hide all columns that were *not* used in the scoring and sorting. The user can selectively unhide certain columns by using the “Edit:GoTo” menu option (or typing “CTRL-G”), enter the columns in the “Reference” box (for example, C:F), then pick the “Format:Column:Unhide” menu option.

10 If the user wants to see a color-score-sorted list of compounds within a particular cluster (such as the best cluster), the user should do the following:

1. Sort by clusters to find the ID of the cluster wanted.
2. With a second rule, sort by compounds.
3. Go to the “SORTED by Compound” results sheet and turn on EXCEL’s “Data:Filter:AutoFilter” feature for the column that specified the 15 clustering in the first sort. The user can then choose to view only the compounds in one particular cluster, and they will be in compound-sorted order.

Example

20 With reference to the already-colored worksheet “DEMO 1” shown in **FIGURES 8A-8B**, the cluster control worksheet shown in **FIGURE 11A**, and the control panel shown in **FIGURE 11D**, application of the scoring and sorting of clusters is described. As noted above, in the control panel of **FIGURE 11D**, the parameters are stored with the name “Cmpd” 1102. The color red has a score of

“.1”, orange has a score of “0”, yellow has a score of “1” and lightgreen has a score of “2”. Columns C and D have relative weights of “1”, as does column E.

Application of control panel “Cmpd” of **FIGURE 11D**, by selecting “Score and Sort Clusters”, produces the worksheets shown in **FIGURES 14A-14B**. When the user selects the “Score and Sort Clusters” button 1108 for the “Cmpd” control panel of **FIGURE 11D**, the system first presents a dialog box (1402 shown in **FIGURE 14A**) asking the user how un-colored cells should be scored for sorting. As noted above, un-colored cells can be scored explicitly by user entries (recommended) or as the average of the colors in the same row. Once the user makes a selection and clicks on the “OK” button, the system scores and sorts the data, producing the display screen shown in **FIGURE 14B**. The system provides a summary of what was done, including the information about the two new sheets (“DEMO 1 SCORES by Cmpd” and “DEMO 1 SORTED by Cmpd un=ze”) which are added to the workbook. **FIGURES 14C-14D** show the data in the newly created worksheet “DEMO 1 SCORES by Cmpd”.

Example

With reference to the already-colored worksheet “DEMO 2” shown in **FIGURES 10A-10B**, the cluster control worksheet shown in **FIGURE 11A**, and the control panel shown in **FIGURE 11C**, application of the scoring and sorting of clusters is described. In the control panel of **FIGURE 11C**, the parameters are stored with the name “acids” (1102). The color red has a score of “0”, orange has a score of “1”, yellow has a score of “2” and light green has a score of “3”. Column D has a relative weight of “1”.

The application of the parameters or rules in the "acids" control panel produces two new worksheets ("DEMO 2 SORTED by acids" and "DEMO 2 SCORES by acids") shown in **FIGURES 15A-15B**.

5

4. Score and Sort Dose-Response Data.

10

15

Data grouping and visualized by color coding has also been found to enable an automated solution to another vexing pattern recognition problem. An HTS lab is currently able to provide dose-response data on some subset of the whole collection of compounds originally tested. Sometimes, logistical constraints (time and/or cost) dictate that only a few concentration points can be run on each compound, and the high-throughput nature of the process generates somewhat noisy data. A similar situation sometimes exists in other biological laboratories where assays are very time-consuming. Dose-response curves with few, noisy points are difficult to analyze by traditional curve-fitting methods. The present invention includes a mechanisms/algorithms for analyzing percent-of-maximal-effect data and accurately ordering the compounds by potency, even when faced with few points and high noise.

20

The mechanism recognizes two properties of the dose-response data for each compound:

1. "Dose-responsiveness," the drop-off of activity with dilution, is taken as a sign that the compound has some reasonable pharmacological mechanism of action.

2. The activity measurements at the various concentrations also provide a confirmation of the general level of each compound's activity that was indicated by the original single-poke HTS hit.

These two properties are somewhat independent, as illustrated by the 5 example of a compound that is 95% active at all tested concentrations. It demonstrates very poor (i.e., no) dose-responsiveness over the range of concentrations tested, but is so active that it should not be ignored, because it might reveal a dose response if tested at even lower concentrations.

By using the data groupings and color codes of the dose-dependent activity 10 data columns, which help to smooth out excessively fine distinctions in the numbers, this invention includes an algorithm to assign numerical scores for dose-responsiveness and overall activity in the dose-response data. Moreover, the algorithm also calculates a smart composite of these two scores, in such a way that a highly active compound will get a high composite score even if its dose-responsiveness is poor. This composite score is capable of extracting useful 15 information, even from very noisy data, and has been validated to correctly order a list of test compounds. The system of this invention adds data columns that report all three scores for each compound, and these columns can themselves be color coded, and thus used in further comparison to other types of data by compound or 20 cluster scoring and sorting as described above.

Moreover, within certain limits, the invention's dose-response scoring algorithm can also be used to make *quantitative* estimates of IC₅₀ values of compounds, even in the presence of large amounts of experimental error. This is accomplished by adding a set of hypothetical marker compounds with known 25 potencies and theoretically calculated activities at the test concentrations. Since

the ordering algorithm is reliable, these markers will be ordered into their appropriate place, and can be used to calibrate the ordering scores in terms of actual IC₅₀'s. In other words, estimates of IC₅₀ for the compounds can be generated by interpolating between the markers in the ordered list of composite scores.

5

Scoring and sorting dose-response data according to the present invention processes several columns of colored dose-response data (activity vs. concentration) to assign three numerical scores that can later also be color coded, and thus used by the "Score and Sort Clusters" mechanism (described above) to compare compounds or clusters of compounds. The three scores are:

10

- (a) degree of dose-responsiveness over the concentration range tested;
- (b) overall activity level demonstrated in the dose-response data; and
- (c) a variably weighted composite of (a) and (b), designed to give high scores for high activity even when dose-responsiveness is poor (e.g., a compound that is highly active at all concentrations).

15

20

The scoring and sorting dose-response data according to the present invention bases its scoring on colors rather than absolute activity numbers. The mechanism takes user input from a *Cluster Control* worksheet, e.g., as shown in **FIGURES 11A-11B**. **FIGURE 11B** shows a control panel from the cluster control worksheet shown in **FIGURE 11A**, wherein the user has selected columns F to I of worksheet "DEMO 1" for scoring dose-response.

The system detects uncolored data, notifies the user, and asks whether to continue. If yes, the system skips the row containing the uncolored data. The system inserts three new columns on the original spreadsheet to contain the new scores, the new columns immediately following the columns of dose-response data. The column headings show the name of the parameter set used for scoring.

5 Preferably, the system offers to regenerate existing table of Sheet Statistics to correct it for newly added score columns. Further, the system offers to sort the data rows by decreasing score. The system also offers to carry out quantitative estimates of IC₅₀ values for the user's compounds, by adding artificial calculated calibration marker compounds.

10

In order to score and sort dose-response data:

- (1) Ensure that the dose-dependent activity data columns are ordered with highest concentration at the left and lowest concentration at the right. To ensure this, the system will remind the user of this requirement and ask him to confirm it when this tool is run. *Note:* If the data are for an undesired effect such as toxicity, the columns should be ordered the opposite way (lowest concentration left, highest right).
- (2) Use the Data Coloring (described above) to color the dose-response data columns.
- (3) Go to the Cluster Control sheet 314 (FIGURES 11A-11D) and pick a control panel 1100 to use. On this sheet, a single control panel extends vertically through the black, blue, red, and green sections, and provides input information for several tools.

5

(4) In the top black section, give this new parameter set a name (1102)

if not already done. The name will be used to label the outputs.

(5) In the blue section, enter a sheet name (1104).

(6) In the red section (1110), enter the colors used to color the dose-dependent data, and relative point scores (1112) to be assigned to these colors.

(7) In the green section (1118), enter the columns which contain the dose-response data (using the same syntax as for Data Coloring).

(8) Click the green-text "Score Dose-Response" button (1120).

10

(9) If the data are expressed as "percent of maximal effect," the user can follow the prompts to add calibration markers and make quantitative estimates of IC₅₀'s.

15

Note on the output of score and sort dose-response: the system inserts three score columns after the dose-dependent data. The three scores are all scaled to a 0-100 range, and have meanings as follows:

(a) degree of dose-responsiveness over the concentration range tested:

20

100 = smoothly decreasing with dilution, spanning the entire range of color groups;

75 = flat dose-response; and

<75 = even more poorly behaved

(b) overall activity level demonstrated in the dose-response data

100 = highest activity color group at all concentrations.

(c) a variably weighted composite of (a) and (b), designed to give high scores for high activity even when dose-responsiveness is poor (e.g. a compound that is highly active at all concentrations).

5

The Dose-Responsiveness Scoring Algorithm

The data columns are ordered left to right, by decreasing concentration.

The scoring algorithm awards positive score points for each dilution step across the data that actually shows a decrease in the activity data *group* (i.e., the color), and to penalize every step that does not. The algorithm uses the following scoring:

- +1 point (awarded) when a dilution step moves to a lower activity group
- 0 points when a dilution step leaves the activity group unchanged
- -3 points (penalty) when a dilution step moves to a higher activity group

The maximum score would go to a compound that shows all the possible color group steps in the right direction, and has no reversals. The minimum score would go to a compound with all the possible reversals, and no correct steps. The program then scales the extremes to 100 and 0, in order to present a consistent interface to the user.

The relative magnitudes of the scoring parameters were empirically arrived at by testing "complete sets" of color patterns. This is possible because of the data simplification afforded by the value grouping. If we define the following numerical parameters:

$C = \text{number of colors used, i.e., number of data value groups}$

$P \equiv$ number of points measured, i.e., number of different concentrations (doses) tested,
 then the entire "universe" of possible color patterns includes (C^P) different cases.
 For some typical values that might be encountered in real HTS data, this total
 5 number of cases is manageable in EXCEL, as shown by TABLE 1, below.

TABLE 1. Total Number of Color Patterns

$P =$ # of conc. Points	$C =$ # of color groups	total number of possible cases
3	3	27
3	4	64
4	3	81
4	4	256
5	3	243
5	4	1024
6	3	729
6	4	4096
7	3	2187
7	4	16384*

* For a spreadsheet with a heading row, this exceeds EXCEL's current limit (for EXCEL 95) by one. This value should not exceed the limit for Excel 97.

10 Scoring was done on several of these complete sets. In each set, the results were sorted by decreasing score and compared to "intuition" for general correctness of ordering of dose-responsiveness, and scanned for cases deemed to be clearly out of order. The (+1, -3) score set was found to produce satisfying ordering, while lesser penalties led to poorly ordered results. More objective tests of ordering (described below) were then used to further validate the algorithm
 15

The case of $P=3$ and $C=3$ is presented below in its entirety for illustration. TABLE 2 (FIGURE 16F) shows artificial data and processing for twenty seven (27) hypothetical compounds. The "percent inhibition" columns represent assay "data." If one defines three groups by breakpoints at 33% and 66%, each cell is

assigned to a data group as shown in the middle set of three columns. Here it is clear that the order of compounds in this table is systematic (111, 112, etc.), to illustrate that the complete set is present. The third set of three columns shows color coding, with the darkest being least active and the lightest being most active.

5 Then the data set was processed by the system to yield dose-responsiveness scores, and the results sorted by this score, giving **TABLE 3** (**FIGURE 16G**), the complete set in order of decreasing dose-responsiveness. **TABLE 3** also shows the intermediate step-scoring and unscaled score points, to aid in following and understanding the algorithm. These points are not displayed
10 by the system itself.

The Overall-Activity Scoring Algorithm

15 The second property of interest to be extracted from the data is the overall activity level exhibited by each compound. As explained above, this is largely independent of the dose-responsiveness.

The data value groups' ordinal index numbers are used as single-point activity measures instead of the original data numbers. Extra weight is given to activity shown at lower concentrations by the simple algorithm of weighting each data column by its serial position, again ignoring the actual concentration values.
20 The scores are then scaled to the range 0 to 100. The results of this scoring on the same complete set are shown in **TABLE 4** (**FIGURE 16H**) which has been re-sorted by decreasing overall activity.

The Composite Scoring Algorithm

Comparison of Tables 3 and 4 (FIGURES 16G & 16H) shows clearly that the compound ordering by dose-responsiveness is quite different from the ordering by overall activity. The user (a chemist) could now color-code the new score 5 columns and use them as independent factors in a larger scoring. However, chemists also want a *single* index of compound quality derived from the dose-dependent data. Moreover, a composite index would further help to alleviate the effects of noise on data interpretation, by incorporating more information into the ordering process. This is an "information-based smoothing" of the data.

10 Therefore, a procedure to calculate a third, "smart composite" score from the other two scores was devised.

The general idea is that when selecting good compounds from dose-response data, compounds showing overall high activity should not be discarded for lack of responsiveness. Therefore, the smart composite score should give 15 more weight to the overall activity when the overall activity is high, but lower weight when it is low. A generalized weighted average is written as

$$\text{composite score} = (\text{activity weight})(\text{activity score}) + \\ (\text{responsiveness weight})(\text{responsiveness score})$$

or, defining corresponding symbols:

$$S_C = (W_A)(S_A) + (W_R)(S_R)$$

If the weights are normalized to sum to unity, then this becomes

$$S_C = (W_A)(S_A) + (1 - W_A)(S_R)$$

The activity weight W_A varies with the activity score S_A in such a way as to achieve the desired result.

The functional form of this variation was the subject of empirical testing. It was decided that the limits would be that W_A would approach 0.5 (activity and responsiveness equally weighted) in the limit of low activity, and that W_A would approach 1.0 (responsiveness ignored) in the limit of high activity. The actual variation was encoded as an exponential increase in order to have rather sharp onset of the activity bias at higher activities:

$$W_A = (C_1) \exp [(k)(S_A)] + C_2$$

The value of the coefficient $k=0.06$, for which the activity bias starts to become substantial around an activity score of eighty (80), was chosen for implementation in a preferred embodiment of this invention, according to empirical results. **FIGURE 16I** shows the variation for a few values of k . **TABLE 5 (FIGURE 16J)** shows all three scores for the example complete set, now sorted by decreasing composite score.

The details of the scoring algorithms were arrived at largely by comparing results to intuitive ordering of the test cases in the complete sets. Because the sets were complete, no really new results can be generated by further test sets. However, one can generate test activity data sets from compounds of known potencies, whose real rank ordering is thus known, in order to see more objectively how well the scoring algorithms rank the results.

To this end, a set of pseudo-ligands was hypothesized, with dissociation constants from a fictitious receptor ranging from nanomolar to millimolar ($pK = 9$ to 3). The set included thirty one (31) compounds, with potencies evenly spaced by 0.2 log units (9.0, 8.8, 8.6, ... , 3.4, 3.2, 3.0). A “pseudo-screen” was created which “tested” binding of these ligands at five concentration points in the usual

range: 10^{-5} M, 3×10^{-6} M, 10^{-6} M, 3×10^{-7} M, and 10^{-7} M. Note that the span of potencies exceeds the span of concentrations tested by two log units on each end, so the test set includes both "very active" and "very inactive" compounds relative to the screening concentrations.

5

Then artificial binding data were created by calculation as follows.

Assuming a simple binding equilibrium of the ligand to a receptor, the "percent inhibition" at a given ligand concentration is equal to the fraction of receptor sites which are occupied by ligand, given by simple equilibrium equations as

10

$$P_{inhib} = 100 \cdot (ligand) / [K + (ligand)]$$

15

For a more realistic simulation, artificial random noise was then added to the calculated numbers. The first experiment reported below used noise randomly distributed over the range of ± 10 inhibition percentage points, and the second with noise up to ± 30 inhibition percentage points. Note that this means ten or thirty percentage points of absolute error, not 10% or 30% of the value.

20

The artificial data were then color-coded according to the mechanisms of this invention (described above) into four color groups, using the simple breakpoints at 25, 50 and 75 percent inhibition. Note that in assigning these breakpoints, no consideration was given to the actual data values. Then the scoring algorithms of this invention were run, and the compounds sorted by the composite score. Rank order numbers were assigned to the compounds, with 1 being the most potent and 31 the least. In cases of ties in the composite score, equal rank numbers were assigned, with a value equal to the average of the rank numbers spanned by the tied group of compounds (e.g., a tie for 2nd and 3rd

25

would result in each compound being ranked "2.5"). For each experiment, the final rankings were plotted against the "real" rankings by known potency, to test how well the scoring algorithms ordered the compounds. These plots are shown in **FIGURE 16K** (for noise=10) and **FIGURE 16L** (for noise=30).

5 For the experiment with noise up to 10 inhibition percentage points, shown in **FIGURE 16K**, the ranking of the composite scores is "perfect" (in the sense of having no inversions) over the range of tested concentrations (pK = 5 to 7). The pseudo-screen is unable to distinguish the potencies of compounds above or below this range.

10 When the noise is much higher (30 percentage points), the ranking of individual compounds is not as precise, but one can identify three cleanly divided "good-medium-bad" groups, as indicated by the dashed boxes on **FIGURE 16L**. Thus, even with this rather extreme noise level, the invention's scoring still successfully prioritizes the compounds into groups. The range where 15 discrimination is effective is still roughly the range of the test concentrations (pK = 5 to 7), but has been reduced somewhat by the higher noise. Note that the ranking within this range (the middle boxed group) is still mostly correct, with only one inversion, even for single compounds.

Quantitative Estimation of Potencies

20 With confidence established that the algorithms provide reliable rankings of compounds by potency, it is possible to proceed to making quantitative estimates. The method uses calibration marker compounds.

To understand this method, it is helpful to realize that the concept is analogous to the quantitative use of SDS polyacrylamide electrophoresis gels to

measure protein molecular weights. The proteins are known to migrate through the gel with speeds directly dependent on molecular weight, but it is difficult to calculate the absolute migration rates for a particular experiment. In dose-response scoring, the compounds are known to be properly ordered, but it is not clear how to calculate a potency (e.g., K_{diss} or IC_{50}) directly from the score.

5 Protein chemists solve the molecular weight problem by running marker proteins, with known molecular weights, in the same gel, then using their band positions as calibration for the unknowns. Analogously, this invention's quantitative estimation method uses hypothetical marker compounds of known 10 potency to internally calibrate the dose-response composite scores for the user's choice of a coloring rule, then interpolates the potencies of the unknowns.

To create markers, the system asks the user to input the concentrations used for each of the dose-dependent activity data columns. The system then picks a set of calibration concentrations, at intervals of 0.5 log units, to span the tested 15 range. For each of these calibration concentrations, a marker compound is created and added to the user's compound list, and artificial data is calculated for each column, from the same simple equilibrium binding equation used above in the validation study (this time with no "noise"):

20
$$p_{inhib} = 100 \cdot (ligand) / [K + (ligand)]$$

The marker data are then colored by the same rule used for the user's compounds, and the scoring and sorting algorithm is re-run.

25 The result is that the markers are sorted into the list according to their potencies, and the potencies of the other compounds can be estimated by

interpolating between the markers, using the composite dose-response scores. To illustrate, a typical section of a sorted list is shown below in **TABLE 6** (**FIGURE 16M**), using four colors.

Potencies for compounds that fall between two markers are calculated by 5 linear interpolation between the logarithms of the markers. Given the various uncertainties in the data values themselves and in the evaluation process, it was found that linear interpolation between markers spaced at 0.5 log unit intervals was sufficiently precise, and no more complex curve fitting was necessary.

Validation of Quantitative Estimation

10 Validation of the quantitative estimation method followed a procedure very similar to that used to validate the scoring, and using the same sets of test data with various noise levels. As before, the testing concentrations were from 10^{-5} to 10^{-7} M (negative log from 5 to 7). Marker compounds (no noise) were added with pK's from 4.5 to 7.5, and K_{diss} estimates for the noisy compounds were 15 carried out by the interpolation method. The results are shown below for the cases of 10 and 30 inhibition percentage points of noise.

FIGURES 16N and 16P show that the estimates are clearly quite good 20 within the range of the testing concentrations (pK 5 to 7), but the quality of estimation deteriorates quickly beyond those limits, and algorithm does not reliably distinguish among compounds whose potencies are more than a half log unit beyond the testing range. Therefore, it was decided that presently preferred embodiments would not report any estimated values that fell outside the range of concentrations used in the testing data columns. Thus, in the example in **TABLE 6** (**FIGURE 16M**), the lowest testing concentration was 10^{-7} M ($= 0.1 \mu\text{M}$). For the

first compound in **TABLE 6**, the system has estimated a potency with $\text{pIC}_{50} > 7$, but it conservatively only reports " $<0.1 \mu\text{M}$."

5 **TABLE 7** summarizes the statistics of the estimations within the testing limits. **TABLE 7** shows that the method successfully estimates the potencies within about a factor of two, even with high noise levels.

TABLE 7. STATISTICS OF ESTIMATION VALIDATIONS

<i>percentage points of noise</i>	<i>number of compounds</i>	<i>average of abs(log error)</i>	<i>Average error in IC_{50} (factor)</i>
10	13	0.22	$\times 1.6$
30	13	0.39	$\times 2.5$

10

Comparison to Other Methods of Quantitative Estimation

15

Further corroboration was obtained by treating some real data from T-cell proliferation blockage assays. It is estimated that these data have *at least* as much noise as the artificial test set with 30 inhibition percentage points added. The standard treatment of this data in the past has been to fit a dose-response curve with a Hill coefficient of 1, using a PC-based program ORIGIN. (ORIGIN is a data analysis program from Microcal Software, Inc. of Northampton, Massachusetts. ORIGIN is used in this instance for non-linear least-squares fitting of dose-response curves to functional equations.)

20

The data used here were from testing in the concentration range from 1 to $0.03 \mu\text{M}$ (negative log from 6 to 7.5). The plot in **FIGURE 16Q** shows the correlation of values estimated by this invention with values from ORIGIN fits.

Two compounds that the present invention estimates to be beyond the testing range, i.e., pIC_{50} below 6, are included as open diamonds, for illustrative purposes explained below. (As explained above, preferred embodiments of this invention would normally not report these values.)

5 The results are consistent with the properties observed in the validation study. The present invention estimates are quite good within the testing range (6 to 7.5). At the lower limit, this invention has made two estimations exactly at 10^{-6} M (arrows) which do not correlate as well with the ORIGIN fits. Nevertheless, because the whole plot spans only a relatively narrow range of potency, even these 10 discrepancies are not very large. For all twenty eight (28) estimates within the testing range (including these two), the average logarithmic deviation between the two estimation methods is 0.18, corresponding to a factor of only 1.5.

It is further noted that the calibration marker estimation method does not uniformly "flatten out" beyond the testing concentration range. The two open 15 diamonds in **FIGURE 16Q** are estimations that presently preferred embodiments of this invention would not normally report because they have $\text{pIC}_{50} < 6$, but they agree well with ORIGIN fits

It is interesting to compare calibration-marker with curve-fitting results for 20 particularly badly behaved data, such as dose-response curves that are not monotonic with respect to concentration. This is sometimes the type of data that emerges from dose-dependent screening in a high-throughput mode. Studies of this type have been initiated by adding artificial noise to the extent of fifty (50) inhibition percentage points.

Finally, it should be pointed out that there is some mechanical advantage 25 to using the present invention relative to current practice of using ORIGIN.

ORIGIN is used by manually filling in a template with data, then manually executing a fit. Depending on the number of points and the degree of customization of parameters, this can take one to ten minutes of the user's time. The present invention, on the other hand, processes a whole spreadsheet at once (i.e., up to 16,383 compounds), and goes at a rate of about 3,000-4,000 compounds per minute on a 200 MHz PC.

Examples

FIGURE 16A shows dose response data for twenty (20) compounds at four concentrations. The data have been grouped and the cells colored by the rule shown in **FIGURE 16B**. The result of the scoring and sorting process is shown in **FIGURE 16C**, where the compounds are ordered by decreasing values of the composite score (column H). Then, virtual "marker" compounds are added with known potencies spaced by 0.5 log units, and they are shown in **FIGURE 16D**, colored by the same rule and scored. The name of each marker compound designates the logarithm of its potency, e.g., "marker_7.0" has a potency $IC_{50} = 10^{-7}$ M. **FIGURE 16E** shows the result of sorting the list by decreasing composite score after adding the markers. This process then enables estimation of IC_{50} values for the compounds by interpolating in the column (H) of ordered composite scores, and these estimates appear in two forms in columns I and J.

6. Summarize Spreadsheet Statistics Mechanism

This mechanism creates a table summarizing the entries in each column of a data sheet, to aid the user in deciding how to color each column. The

mechanism counts numeric, text, and data entries, and uses color to flag columns that have mixed types. The mechanism also counts blanks, and specially flags columns with “trailing blanks,” i.e., columns which are shorter than the longest one on the spreadsheet. For numeric data, the mechanism calculates minimum, maximum, mean, and standard deviation, even in the presence of interspersed text entries. For text data, the mechanism presents a list of the text strings used and their occurrence counts. The mechanism creates a summary key of the column letters and headings as a text box that can be copied to other sheets for convenient reference.

FIGURE 17A shows a sample spreadsheet containing miscellaneous data on twenty four (24) compounds. **FIGURE 17B** is the statistics sheet calculated from it. Each row of the statistics sheet describes one column of the original data sheet. First, the counts of numeric, text, date, and blank entries are listed, followed by two columns describing the total length of the data sheet. Then the minimum, maximum, mean, and standard deviation of any numeric data are reported. Finally, the statistics sheet lists a summary of the text strings found in each original data column. As examples, in **FIGURE 17B**, one can see that original column A (“Cmpd”) had twenty four (24) different text strings, that the numeric data in original column C (“Test1”) had a mean of 2.385, and, flagged by the red coloring, that original data column E (“Test3”) had a mixture of ten numeric data and two text strings, both “N.A.”

The details of how the program code accomplishes this are straightforward, and one of ordinary skill in the art would know, from this description, including the Figures, how to make and use this invention. The

program loops through all the entries in the column, testing the data type of each, and tallying the counts and numerical statistics.

Spreadsheet Creation and Organization

5 The operations of this invention require a considerable amount of user input, e.g., to create well-structured spreadsheets, to define and apply diverse coloring rules for large numbers of columns, and to use these colors and the user's stated scientific priorities to create meaningfully ordered lists of compounds or clusters.

10 The user interface of this invention has been designed to ease this process and help the scientist focus on the tasks of formulating and recording clear descriptions of the evaluation parameters. Accordingly, this invention provides a number of tools and mechanisms to aid in the creation and organization of spreadsheets. These tools and mechanisms include:

15

- Smart Append Column Mechanism
- Merge Data Mechanism
- Data Import Mechanism
- Workbook Navigation Shortcuts
- Conversion of "uM" to μM and "UU" to μ
- Delete Pictures Mechanism
- Change Values in Column Mechanism
- Concatenate Values across Columns Mechanism
- Delete Leading Inequality Signs Mechanism
- Delete Derived Sheets Mechanism

20

Smart Append Column

This mechanism appends new columns of data onto an existing spreadsheet, matching rows by labels (e.g., compound numbers). The mechanism 5 copies all data to a new sheet before doing its work, leaving the original sheets unchanged. There is no need for the user to pre-sort any of the data. The mechanism provides optional case-sensitive or case-insensitive label matching.

New rows are added at the bottom when new labels do not match any old labels. Rows with missing labels are identified and the system offers to fill them 10 by copying previous label. Rows with repeated labels (i.e., replicate data) are also identified and the system offers a choice from among several automated processing rules, or manual fixing. A fast matching algorithm temporarily sorts rows by label, then restores original order when finished. Several intermediate stopping points are offered and extra data viewing options for conservative users 15 worried about errors.

Merge Data Mechanism

The merge data mechanism copies new data values from an appended column into an older column. The mechanism copies all data to a new sheet 20 before doing its work, leaving the original sheets unchanged. The mechanism detects cells where new data would overwrite old data that is different, flags them with color, and alerts the user. Several intermediate stopping points are offered to

the user, as are extra data viewing options are offered for conservative users worried about errors.

Data Import

5 One-button (or one-menu-click) import of existing EXCEL spreadsheets into an integrated file, which contains both the data and the related control sheets. The mechanism offers to search for and remove any leading or trailing spaces in the imported data and offers to consolidate replicate data rows into unique ones, using user choices as to how to handle the replicate data. The mechanism also
10 detects hidden rows and offers to unhide them and detects formulas and offers to convert them to values. This mechanism is also used to update to newer version of the system.

Workbook Navigation Shortcuts

15 The system includes various workbook navigation shortcuts including:
• A special added drop-down menu which includes commands for jumping directly to the various control sheets. These commands also have keyboard shortcuts assigned to them.
• From a cell on a control sheet that contains the name of a data sheet, a special item on the right-click shortcut menu jumps directly to that data sheet. Other special items on this menu enable a “Twin Screen” display to
20 see two sheets at once.

- To aid in choosing columns to enter on control sheets, there is a special “Twin Screen” display triggered by right-clicking any cell on a control sheet that contains the name of a data sheet.

5

Convert “uM” to μ M and “UU” to μ

10

Preferred embodiments of the system of this invention require the data spreadsheet to have *one and only one* row of column headings. The user can type either of the encoded strings “uM” (lowercase u, uppercase M) or “UU” (both uppercase) into any column heading, select the cell or whole row of headings, then pick this command. Each “uM” in the selection will be converted to “ μ M”, and each “UU” will be converted to a “ μ ”. The code recognizes the special exception of the word “VACUUM” as long as it doesn’t end with the cases “uM.” This conversion allows the user to avoid the confusing use of lowercase “u” or the column-widening use of the full prefix “micro.” This utility appears on the system menu.

15

Delete Pictures

20

The system provides a mechanism for removing pictures containing chemical structures, in order to reduce file size, processing time, and confusion when they do not align properly after row sorting.

Change Values in Column

This is a mechanism for regularizing data in a spreadsheet column. It facilitates replacement of all occurrences of a given value by another. The mechanism creates backup copies of the original column, and updates any existing data statistics for the edited sheet.

5

Concatenate Values across Columns

The system provides a mechanism for regularizing data in a spreadsheet column. Some possible uses include: (a) construction of unique row labels: M-number plus stroke number → “M123456/001”; and (b) reconstitution of numerical inequalities from separate columns: “>” plus a number → “>number”. The user is provided with an option to include linking (delimiting) text strings between values and an option to include or skip blanks. The system retains the original columns and inserts a new one for the results.

10
15

Delete Leading Inequality Signs

Another mechanism for regularizing data in a spreadsheet column includes the mechanism to delete leading inequality signs. This mechanism converts entries like “>1000” to just the number “1000”. This must be used with considerable caution, because it is the equivalent of creating a false test result. It is generally preferable to color the cells containing text strings with the data coloring mechanism described above, rather than alter them. All later processing is based on the colors, not the cell values. This mechanism also deletes inequality

20

sign only if it is the first character in the cell. The mechanism creates backup copies of the original columns.

Delete Derived Sheets

5 The system menu includes a command to delete all output sheets from the current workbook, with separate user confirmation for each one. This is intended as a cleanup mechanism for information that may be outdated and is easily regenerated by subsequent system runs.

10 Initial experience with the coloring tool has revealed that color coding has more subtle, but far-reaching usefulness. The colors themselves also can act as a “currency of exchange,” a medium for comparing the quality of one kind of result to the quality of a very different kind of result. For example, an HTS activity of “95% inhibition” may be considered desirable and color coded, e.g., green. In the 15 same list of compounds, a molecular weight between 400 and 600 may be considered optimally desirable, and thus also color-coded green. If the user takes care when assigning colors, “green” takes on a common meaning across the board. This translation of data values into colors then opens up a cornucopia of possibilities for processing the colors (as numerical color indices) and comparing compounds, searching, in our example, for the ones that are the “most green.” 20

Accordingly, the system includes tools to numerically score individual compounds or clusters of compounds by the colors that appear in their various data columns. The system can then create a new spreadsheet sorted by this score

(either by single compounds, or cluster-by-cluster, the choice being the user's), in which the "most green" compounds will then appear at the top.

5

Examples

Application to Portfolio Management

The system can equally well be applied to any set of data where the rows are cases of a similar construct, with the columns being various properties of each case. For example, a data spreadsheet can contain a list of competing projects or investments for a company's portfolio, with the columns containing various managers' ratings of each project or investment. **FIGURE 18A** shows an example of twenty projects, each of which has been scored 1, 2, or 3 on two factors, one more important than the other, by each of three managers. The sheet has been colored by the rule shown in **FIGURE 18B**. Then the data were scored and sorted by the sorting rule of **FIGURE 18C**, and the result is shown in **FIGURE 18D**.

Clearly, the projects that were given a "3" in the important factor come to the top, and it can be seen that the less important factor does indeed matter less to the final ordering. The colors also help to flag anomalies, such as a low score by one manager on an otherwise high-ranking project.

20

In general, the data can be various sorts of data. Some examples are listed below and illustrated in the referenced Figures.

FIGURE 19 shows a list of drug candidate compounds, scored and sorted by a composite of ten parameters that describe their physical, chemical, and biological properties. Green shades indicated desirable values; red shades are

undesirable. The display is compressed vertically with the vertical re-scaling tool to clearly display the difference in coloring patterns between the top eighty (80) compounds and the bottom eighty (80) compounds (separated in the illustration by a blank band).

5 **FIGURE 20** shows a list of proteins that are candidates for targets for drugs, chosen from a pool of candidate genes, scored and sorted by a composite of eleven parameters that describe their suitability.

10 **FIGURE 21** shows a list of research projects competing for resources. Each project has been scored according to several evaluation factors, and the whole array has been sorted by color groups. The same construct is useful for evaluating 15 employee performance or job candidates.

15 **FIGURE 22** shows a list of pharmaceutical companies and their current status with regard to discovering or marketing products in each of various disease areas. Each company's line of products has been scored according to the maturity of the offerings, and the whole array has been sorted by color groups.

20 **FIGURE 23** shows the use of data-grouping (coloring) rules to visualize the time courses of drug concentrations in blood. In this example, light colors were chosen to represent high concentrations of drug in the blood, while dark colors were chosen to represent low concentrations. The figure shows a wide range of differing time courses.

FIGURE 24 shows the use of data-grouping (coloring) rules to visualize the matrix of pairwise cross-correlations of the results of eight (80) drug screens. In this example, light colors were chosen to represent low correlations, while dark colors represent high correlations.

Quantitation of the Similarity of Data Grouping in Two Variables

As part of the present invention, a mechanism is provided for assigning a quantitative measure to the degree of similarity of grouping (visualized by color coding) of data in each of two columns of an EXCEL spreadsheet. The mechanism allows for a correlation-like analysis on a wide variety of data types, including text, or mixed numbers and text.

In the data-exploration paradigm of the present invention, one of the first steps a user takes is to divide the range of data values in each column into a small number of groups for further analysis, thus effecting a reduction of precision which has been found to be useful in various ways.

It is sometimes useful to explore whether the rows of the data matrix have been divided into similar groupings in each of two different columns. For example, a researcher might ask, "Do the high molecular weight compounds tend to be the ones whose solubilities fall below the limits of measurement?" In other words, this would mean to compare the groupings in the molecular weight column with the groupings in the solubility column.

If the data were strictly quantitative, this would be called *correlation* of variables, and there exist a number of perfectly good statistical measures of the phenomenon. However, one of the unique capabilities of the present invention lies in dealing with textual data and mixtures of numbers and text, and it would be helpful if one could translate the visible color patterns of the present invention to some kind of quantitative measure of correlation. In order to avoid confusion with standard statistical correlation, the distinct term "color grouping similarity" is used to describe the new measure.

In preferred embodiments of this invention, the data grouping is stored in the form of the colored backgrounds of data cells. At first glance, one might simply seek to compare the colors of the first column with those of the second, and count the number of rows with matching colors. However, a color grouping similarity tool must be able to cope with the possibility that the colors are different. This could happen because the user chooses completely *different color schemes* for the two columns, or because the correlation is negative. As an example of negative correlation, suppose column A contains random numbers between 0 and 1, colored such that those above 0.5 are green and those below 0.5 are red. Then imagine a column B where each value is equal to one minus the corresponding value in column A, i.e., the “one’s complement.” If the user colors the second column with exactly the same coloring rule as the first, every row will have a different color in column B than in column A. None of the colors will match, though the groupings are perfectly correlated. To be successful, the tool must deliver a high measure of correlation between such pairs of columns. The algorithm described below was designed to perform in this way.

Algorithm for Measuring Data Grouping Similarity

The algorithm was derived from semi-quantitative reasoning, as follows. It is based on the qualitative question, “For all rows that have one particular color in the first column, to what degree do they have a *uniform* color in the second column (not necessarily the *same* color as in the first column)?” The quantified answer to this question is then averaged over the set of colors used.

The details of the mechanism can be seen by example. First, to compare the grouping in two columns A and B, a matrix of “ordered color pair counts” (OCPC) is defined such that each matrix element $OCPC_{ij}$ is the count of spreadsheet rows where one finds color i in spreadsheet column A and color j in spreadsheet column B. Then, the rows of the two spreadsheet columns are scanned to count the number of occurrences of each ordered color pair and thus to determine the values of the matrix elements.

5

In the discussion below, carefully distinguish the rows and columns of the user’s *spreadsheet* of data from the rows and columns of the derived OCPC *matrix*.

10

If the two spreadsheet columns had *exactly the same coloring*, the nonzero elements of the OCPC matrix would all be on the diagonal. As a simplified example, consider four colors (green, yellow, orange, red) and a total of 16 data rows. The diagonal matrix might be (zero elements left blank for emphasis)

15

color in column A	color in column B →	green	yellow	orange	red
green		4			
yellow			2		
orange				7	
red					3

In the case above, there are groups of 4 spreadsheet rows colored green (in both columns), 2 rows colored yellow, 7 rows colored orange, and 3 rows colored red.

20

In contrast, if the groupings were the same, but the coloring rule for the second spreadsheet column used the same colors in a different order, the OCPC matrix might look like the following, no longer diagonal:

color in column A	color in column B →	green	yellow	orange	red
green				4	
yellow		2			
orange					7
red			3		

A simple extension applies if *entirely different colors* (cyan, blue, maroon, purple) are used in the second spreadsheet column. The OCPC matrix might then
5 be:

color in column A	color in column B →	cyan	blue	maroon	purple
green				4	
yellow		2			
orange					7
red			3		

In *any* of the three cases above, the groupings are identical, and the OCPC
10 matrices have the property that each matrix row and matrix column has only one
nonzero element. That lone element is necessarily equal to the sum of the row or
column. This situation should receive the *highest* similarity score.

One way to define the contrasting situation that would deserve the *lowest*
similarity score would be that for each user-defined group of spreadsheet rows in
one spreadsheet column, the other spreadsheet column has a “maximally non-
unique” set of colors. In the corresponding OCPC matrix, this corresponds to
each matrix row or column having a broad distribution of values rather than a
single non-zero, a uniform distribution has been chosen as the definition of this
state:

color in column A	color in column B →	cyan	blue	maroon	purple
green		1	1	1	1
yellow		1	1	1	1
orange		1	1	1	1
red		1	1	1	1

This low-similarity state can be more precisely defined by saying that each element in a given matrix row or column is the average of all the counts in that matrix row or column.

5 With these concepts defined and the OCPC matrix filled, the scores can then be derived. Each OCPC matrix row (corresponding to a color group in the first spreadsheet column) is selected in turn for scoring. Each element in the matrix row is given a score between zero and one, according to its linear interpolation between: on the one extreme, the average of the nonzero elements in 10 the row, and on the other, the sum of the row or column (i.e., the maximum value it could have if all the others were zero). The scores are then averaged over all the rows of the OCPC matrix to generate a row-wise score component.

15 Next, the corresponding process is applied to the *columns* of the OCPC matrix (each corresponding to a color group in the second spreadsheet column rather than the first). The resulting column-wise score component is averaged with the row-wise score component, then the average is scaled to a maximum of 100 to generate the final similarity score for the two spreadsheet columns.

Interpretation of the Similarity Scores

20 Although the scores are quantitative and well-defined, their interpretation is best done in a partly subjective manner, based on experience. The behavior of

the scores is best understood by example. In **FIGURE 25**, the leftmost (“base”) column has been compared to each of the others, and the scores are shown as well as pictures of the grouping patterns. Comparison of the base with the next column shows that the tool delivers a maximal score of 100 for identical grouping, even 5 when the colors are completely different. Then, stepping across the figure toward the right, it can be seen how the score decreases as the grouping pattern gradually becomes less similar to that of the base column. All the way down to a similarity score of 40, it is still basically true that the light colors are on top and the dark on the bottom, with increasing “noise,” but when the score falls to 20, the pattern 10 appears to have no correspondence to that of the base.

Implementation of the Data Grouping Similarity Tool

In practice, the tool allows the user to choose two sets of spreadsheet data 15 columns. The program then automatically generates all pairs containing a column from the first set with a column from the second set, then writes the similarity scores onto a newly inserted spreadsheet in the user’s workbook. The output takes the form of a table where the degree of similarity is itself color-coded to aid the user in identifying significant cases. An example appears **FIGURE 26**.

20

While the invention has been described with reference to particular mechanisms (algorithms, processes and functions) and architectures, one skilled in

the art would realize that other mechanisms and/or architectures could be used while still achieving the invention.

While embodiments of the present invention have been described with particular setup and initialization procedures, other setup and/or initialization procedures can be used.

Further, while many of the operations have been shown as being performed in a particular order, one skilled in the art would realize that other orders, including some parallelization of operations, are possible and are considered to be within the scope of the invention.

While the present invention has been described with reference to analysis and pattern recognition in data sets relating to chemical compounds, the methods, systems and devices of this invention are considered to be general constructs covering other, non-chemical data sets.

Thus, are provided methods, systems and devices for analysis and pattern recognition in large, multidimensional data sets using low-resolution data grouping. One skilled in the art will appreciate that the present invention can be practiced by other than the described embodiments, which are presented for purposes of illustration and not limitation, and the present invention is limited only by the claims that follow.

What is claimed is:

1. A method of operating on data, the method comprising:
providing at least one user-defined grouping rule for grouping the data into
5 a user-definable number of groups; and
applying at least one of the grouping rules to the data.
2. A method as in claim 1 wherein the data are provided in a table and
wherein the at least one grouping rule applies to at least one user-selectable
10 column of the table.
3. A method as in claim 1 wherein the at least one grouping rule
defines breakpoints corresponding to the user-definable number of groups, and
wherein application of the at least one rule to the data divides the data into groups
15 based on the breakpoints.
4. A method as in claim 1 further comprising:
presenting the grouped data in a manner that visually distinguishes the
groups.
20
5. A method as in claim 4 wherein the grouping rules associate colors
with groups and wherein the presenting of the grouped data further comprises:
coloring an aspect of the data according to the rules.

6. A method as in claim 4, wherein the data are in labeled columns in a spreadsheet, and wherein the at least one grouping rule specifies at least one breakpoint and a corresponding color for each at least one breakpoint, and wherein the presenting of the grouped data comprises:

5 coloring each data item in the at least one labeled column of the data based on the at least one breakpoint and the corresponding color of the at least one breakpoint.

7. A method as in any one of claims 3 and 6, wherein the breakpoints
10 are selected from: (a) numeric values; and (b) textual values.

8. A method as in claim 3 wherein the at least one breakpoint is determined automatically based on the data.

15 9. A method as in claim 5 wherein the data are provided in a table, wherein the coloring of an aspect of the data comprises:

coloring backgrounds of table cells according to the rules.

20 10. A method as in claim 1 wherein the number of groups is fewer than a number of possible data values.

11. A method of operating on data, the method comprising:
providing at least one user-defined grouping rule for grouping the data into a user-definable number of groups;

applying at least one of the grouping rules to the data to generate grouped data;

providing at least one user-defined scoring rule for scoring the grouped data according to user-defined scores; and

5 applying at least one of the scoring rules to the grouped data to score the grouped data.

12. A method of operating on data, the method comprising:
generating grouped data by applying to the data at least one user-defined grouping rule for grouping the data into a user-definable number of groups; and
10 scoring the grouped data by applying to the grouped data at least one user-defined scoring rule for scoring the grouped data according to user-defined scores.

13. A method according to claim 11 or 12 wherein the data comprises a number of parameters for each of a number of cases and the scoring rule 15 comprises a scoring function of user-selectable parameters and user-defined weights for the selected parameters to be used in scoring the cases, wherein the scoring of the grouped data comprises:
scoring the function to the data to obtain a score for each case.

20
14. A method according to claim 13, further comprising:
sorting the scored cases by score.

15. A method according to claim 14, wherein the scored cases are 25 sorted individually.

16. A method according to claim 14, wherein the scored cases are sorted by cluster.

5 17. A system for operating on data, the system comprising:
a mechanism constructed and adapted to provide at least one user-defined grouping rule for grouping the data into a user-definable number of groups; and
a mechanism constructed and adapted to apply at least one of the grouping rules to the data.

10 18. A system as in claim 17 wherein the data are provided in a table and wherein the at least one grouping rule applies to at least one user-selectable column of the table.

15 19. A system as in claim 17, wherein the at least one grouping rule defines breakpoints corresponding to the user-definable number of groups, and wherein application of the at least one rule to the data divides the data into groups based on the breakpoints.

20 20. A system as in claim 17, further comprising:
a mechanism constructed and adapted to present the grouped data in a manner that visually distinguishes the groups.

21. A system as in claim 20, wherein the grouping rules associate colors with groups and wherein the mechanism constructed and adapted to present the grouped data further comprises:

5 a mechanism constructed and adapted to color an aspect of the data according to the rules.

22. A system as in claim 20, wherein the data are in labeled columns in a spreadsheet, and wherein the at least one grouping rule specifies at least one breakpoint and a corresponding color for each at least one breakpoint, and wherein the mechanism constructed and adapted to present the grouped data comprises:

10 a mechanism constructed and adapted to color each data item in the at least one labeled column of the data based on the at least one breakpoint and the corresponding color of the at least one breakpoint.

15 23. A system as in any one of claims 19 and 22, wherein the breakpoints are selected from: (a) numeric values; and (b) textual values.

24. A system as in claim 19 further comprising:
20 a mechanism constructed and adapted to determine at least one breakpoint automatically, based on the data.

25. A system as in claim 21 wherein the data are provided in a table, wherein the mechanism constructed and adapted to color an aspect of the data comprises:

a mechanism constructed and adapted to color backgrounds of table cells according to the rules.

26. A system as in claim 17 wherein the number of groups is fewer
5 than a number of possible data values.

27. A system of operating on data, the system comprising:
a mechanism constructed and adapted to provide at least one user-defined grouping rule for grouping the data into a user-definable number of groups;
10 a mechanism constructed and adapted to apply at least one of the grouping rules to the data to generate grouped data;
a mechanism constructed and adapted to provide at least one user-defined scoring rule for scoring the grouped data according to user-defined scores; and
a mechanism constructed and adapted to apply at least one of the scoring
15 rules to the grouped data to score the grouped data.

28. A system of operating on data, the system comprising:
a mechanism constructed and adapted to generate grouped data by
applying to the data at least one user-defined grouping rule for grouping the data
20 into a user-definable number of groups; and
a mechanism constructed and adapted to score the grouped data by
applying to the grouped data at least one user-defined scoring rule for scoring the
grouped data according to user-defined scores.

29. A system according to claim 27 or 28 wherein the data comprises a number of parameters for each of a number of cases and the scoring rule

comprises a scoring function of user-selectable parameters and user-defined

weights for the selected parameters to be used in scoring the cases, wherein the

mechanism constructed and adapted to score of the grouped data comprises:

5 a mechanism constructed and adapted to apply the function to the data to obtain a score for each case.

30. A system according to claim 29, further comprising:

10 a mechanism constructed and adapted to sort the scored cases by score.

31. A system according to claim 30, wherein the scored cases are

sorted individually.

15 32. A system according to claim 30, wherein the scored cases are

sorted by cluster.

33. A computer-readable memory medium encoded with program data

representing a computer program that can cause a computer to implement a

20 method of operating on data, the method comprising:

providing at least one user-defined grouping rule for grouping the data into a user-definable number of groups; and

applying at least one of the grouping rules to the data.

34. A medium as in claim 33 wherein the data are provided in a table and wherein the at least one grouping rule applies to at least one user-selectable column of the table.

5 35. A medium as in claim 33 wherein the at least one grouping rule defines breakpoints corresponding to the user-definable number of groups, and wherein application of the at least one rule to the data divides the data into groups based on the breakpoints.

10 36. A medium as in claim 33, wherein the method further comprises: presenting the grouped data in a manner that visually distinguishes the groups.

15 37. A medium as in claim 36 wherein the grouping rules associate colors with groups and wherein the presenting of the grouped data further comprises: coloring an aspect of the data according to the rules.

20 38. A medium as in claim 36, wherein the data are in labeled columns in a spreadsheet, and wherein the at least one grouping rule specifies at least one breakpoint and a corresponding color for each at least one breakpoint, and wherein the presenting of the grouped data comprises:

25 coloring each data item in the at least one labeled column of the data based on the at least one breakpoint and the corresponding color of the at least one breakpoint.

39. A medium as in any one of claims 35 and 38, wherein the breakpoints are selected from: (a) numeric values; and (b) textual values.

5 40. A medium as in claim 35 wherein the at least one breakpoint is determined automatically based on the data.

10 41. A medium as in claim 37 wherein the data are provided in a table, wherein the coloring of an aspect of the data comprises:
coloring backgrounds of table cells according to the rules.

42. A medium as in claim 33 wherein the number of groups is fewer than a number of possible data values.

15 43. A computer-readable memory medium encoded with program data representing a computer program that can cause a computer to implement a method of operating on data, the method comprising:

providing at least one user-defined grouping rule for grouping the data into a user-definable number of groups;

20 applying at least one of the grouping rules to the data to generate grouped data;

providing at least one user-defined scoring rule for scoring the grouped data according to user-defined scores; and

25 applying at least one of the scoring rules to the grouped data to score the grouped data.

44. A computer-readable memory medium encoded with program data representing a computer program that can cause a computer to implement a method of operating on data, the method comprising:

5 generating grouped data by applying to the data at least one user-defined grouping rule for grouping the data into a user-definable number of groups; and scoring the grouped data by applying to the grouped data at least one user-defined scoring rule for scoring the grouped data according to user-defined scores.

10 45. A medium according to claim 43 or 44, wherein the data comprises a number of parameters for each of a number of cases and the scoring rule comprises a scoring function of user-selectable parameters and user-defined weights for the selected parameters to be used in scoring the cases, wherein the scoring of the grouped data comprises:

15 applying the function to the data to obtain a score for each case.

46. A medium according to claim 44, the method further comprising: sorting the scored cases by score.

20 47. A medium according to claim 46, wherein the scored cases are sorted individually.

48. A medium according to claim 46, wherein the scored cases are sorted by cluster.

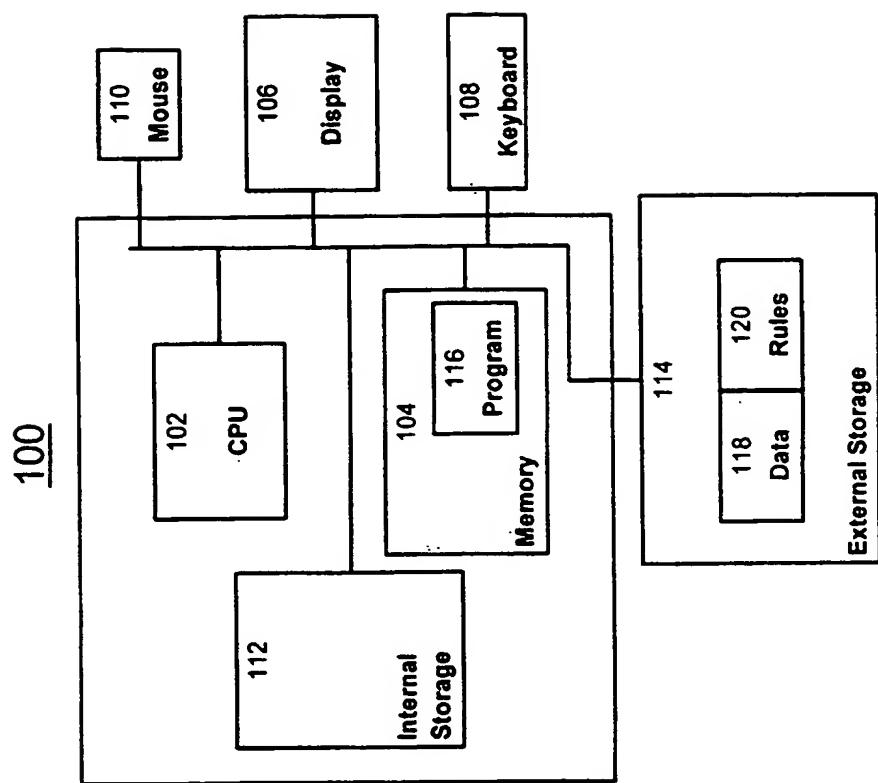
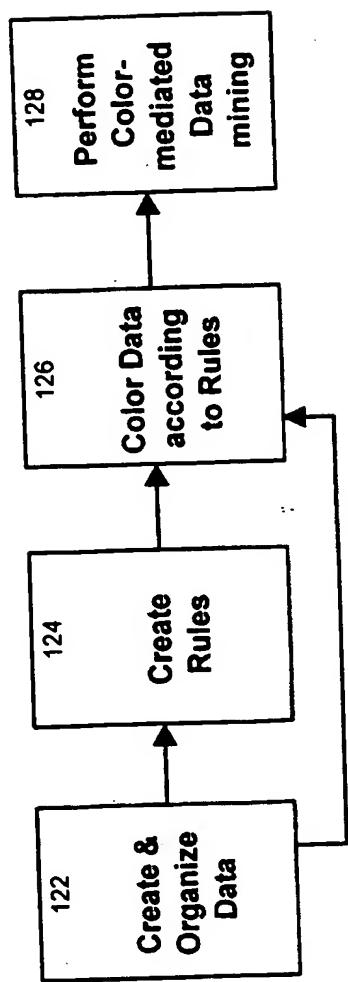


Fig. 1

Fig. 2



AI		Cmpd		HTS SPA		HTS SPA		HTS SPA		HTS SPA	
		Series	Test1	Test2	Test3	Dose-Resp % Inhib @ 3x10-6M	Dose-Resp % Inhib @ 1x10-6M	Dose-Resp % Inhib @ 3x10-7M	Dose-Resp % Inhib @ 1x10-7M	Dose-Resp % Inhib @ 3x10-8M	Dose-Resp % Inhib @ 1x10-8M
1	Cmpd01	N		29	30	41	3	22	5		
2	Cmpd02	N		42	55			28	15		
3	Cmpd03	C		261	11	83	57	24	29		
4	Cmpd04	N		30	70	25	24	21	22		
5	Cmpd05	N	1.8	92	69	60	41	13	3		
6	Cmpd06	D	8.86	6.5	37	71	79	48	43		
7	Cmpd07	D	3.11	0.037	7.8	100	28	28	38		
8	Cmpd08	D	0.089	26	65	68	41	22	15		
9	Cmpd09	D	0.119			61	42	24	5		
10	Cmpd10	N	0.233			50	77	63	25		
11	Cmpd11	N	4.31			47	25	24	3		
12	Cmpd12	H	1.3	0.24		81	59	40	37		
13	Cmpd13	H	1.17	0.194	30	39	23	4	12		
14	Cmpd14	H	0.26	0.41		101	82	38	18		
15	Cmpd15	H	0.369	0.148		99	46	46	36		
16	Cmpd16	I		0.87	30	81	64	47	24		
17	Cmpd17	K		0.223	30	79	54	22	32		
18	Cmpd18	I	5.27			71	71	23	12		
19	Cmpd19	I	0.134			101	109	108	100		
20	Cmpd20	F	0.317			87	70	31	13		
21	Cmpd21	K	221			94	77	36	12		
22	Cmpd22	B		0.15		96	61	36	12		
23	Cmpd23	B				110	91	69	29		
24	Cmpd24	B		0.27		105	104	75	32		
25	Cmpd25	B	0.041	1.1		93	71	41	22		
26	Cmpd26	B	0.665			97	73	43	23		
27	Cmpd27	B	0.111			95	93	52	25		
28	Cmpd28	E		0.13		68	62	12	11		
29	Cmpd29	J		0.46		41	48	9	17		
30	Cmpd30	J	2.79	45							

300

Fig. 3A

312 308 306 304 302

310 308

306 304 302

314

Fig. 3B

300

File	Edit	View	New	Open	Print	Search	Help	Exit
Row	Compd	Series	Test1	Test2	Test3	Test4	Test5	Test6
33	Cmpd34	J	227	24	30	105	62	21
34	Cmpd35	J	0.82			4	28	18
35	Cmpd36	J		0.23		93	63	3
36	Cmpd37	N	12.3			68	20	42
37	Cmpd38	L	0.009	0.024		90	60	14
38	Cmpd39	F	0.06	0.41		87	73	2
39	Cmpd40	G	0.068	1.04		65	66	15
40	Cmpd41	G	0.018	0.35		66	46	40
41	Cmpd42	F	0.38	0.9		102	87	11
42	Cmpd43	H	4.5	0.13		3	25	28
43	Cmpd44	G	0.076	0.1		78	60	13
44	Cmpd45	F	0.5			111	110	92
45	Cmpd46	M	1.14			25	21	21
46	Cmpd47	F	0.026			109	104	97
47	Cmpd48	F	0.27			100	102	80
48	Cmpd49	M	0.036	0.043		71	43	75
49	Cmpd50	F	0.051			112	111	31
50	Cmpd51	G	0.079	0.117		78	70	16
51	Cmpd52	H	0.33	4.9		69	24	27
52	Cmpd53	A	0.036	0.18	16	108	102	23
53	Cmpd54	A	0.33			91	78	14
54	Cmpd55	G	9.12			61	64	27
55	Cmpd56	C	0.39			93	80	33
56	Cmpd57	A	0.018			101	72	60
57	Cmpd58	C	0.22			92	69	26
58								37
59								49
60								49
61								310
62								308
63								306
64								304
65								302
66								312
67								314
68								314
69								314
70								314
71								314
72								314
73								314
74								314
75								314
76								314
77								314
78								314
79								314
80								314
81								314
82								314
83								314
84								314
85								314
86								314
87								314
88								314
89								314
90								314
91								314
92								314
93								314
94								314
95								314
96								314
97								314
98								314
99								314
100								314
101								314
102								314
103								314
104								314
105								314
106								314
107								314
108								314
109								314
110								314
111								314
112								314
113								314
114								314
115								314
116								314
117								314
118								314
119								314
120								314
121								314
122								314
123								314
124								314
125								314
126								314
127								314
128								314
129								314
130								314
131								314
132								314
133								314
134								314
135								314
136								314
137								314
138								314
139								314
140								314
141								314
142								314
143								314
144								314
145								314
146								314
147								314
148								314
149								314
150								314
151								314
152								314
153								314
154								314
155								314
156								314
157								314
158								314
159								314
160								314
161								314
162								314
163								314
164								314
165								314
166								314
167								314
168								314
169								314
170								314
171								314
172								314
173								314
174								314
175								314
176								314
177								314
178								314
179								314
180								314
181								314
182								314
183								314
184								314
185								314
186								314
187								314
188								314
189								314
190								314
191								314
192								314
193								314
194								314
195								314
196								314
197								314
198								314
199								314
200								314
201								314
202								314
203								314
204								314
205								314
206								314
207								314
208								314
209								314
210								314
211								314
212								314
213								314
214								314
215								314
216								314
217								314
218								314
219								314
220								314
221								314
222								314
223								314
224								314
225								314
226								314
227								314
228								314
229								314
230								314
231								314
232								314
233								314
234								314
235								314
236								314
237								314
238								314
239								314
240								314
241								314
242								314
243								314
244								314
245								314
246								314
247								314
248								314

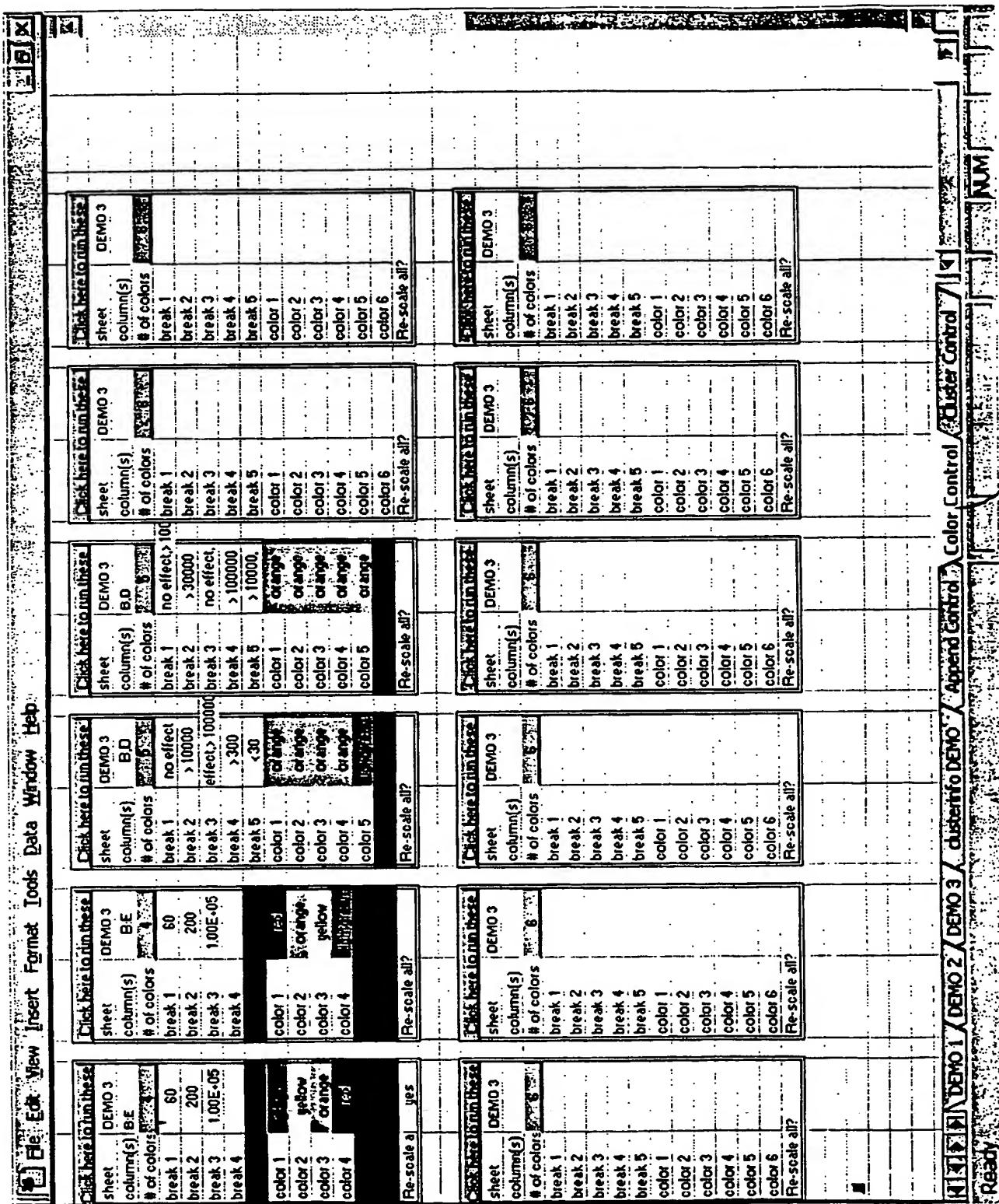
$$130-1 = 140 - 130-2 = 132 - 134 = 136 - 138 = 142 - 144$$

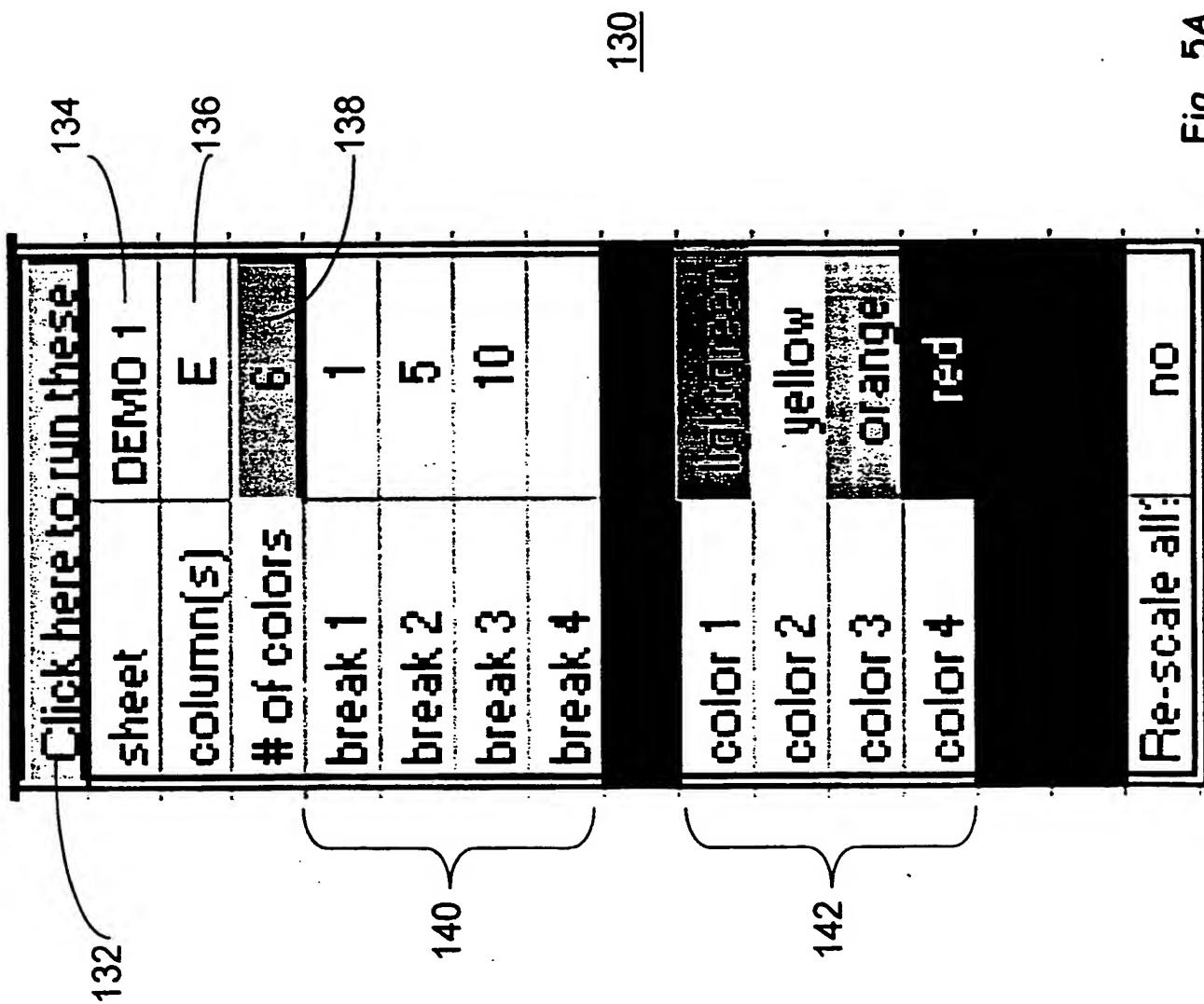
१६

8
11
3
4

Fig. 4A

Fig. 4B





130-1

Click here to run these	
sheet	DEMO 1
column(s)	C,D
# of colors	4
break 1	0.1
break 2	1
break 3	5
break 4	
color 1	
color 2	yellow
color 3	orange
color 4	red
Re-scale all: <input type="checkbox"/> no	

Fig. 5B

The diagram shows a table with various cells highlighted by numbered callouts:

- Callout 152 points to the bottom-left corner of the table.
- Callout 144 points to the top-left corner of the table.
- Callout 146 points to the cell containing "sheet".
- Callout 148 points to the cell containing "color 1".
- Callout 150 points to the cell containing "color 6".

There is no base	
sheet	DEMO 3
column(s)	6
# of colors	6
break 1	
break 2	
break 3	
break 4	
break 5	
color 1	
color 2	
color 3	
color 4	
color 5	
color 6	
Re-scale all?	

Fig. 6A

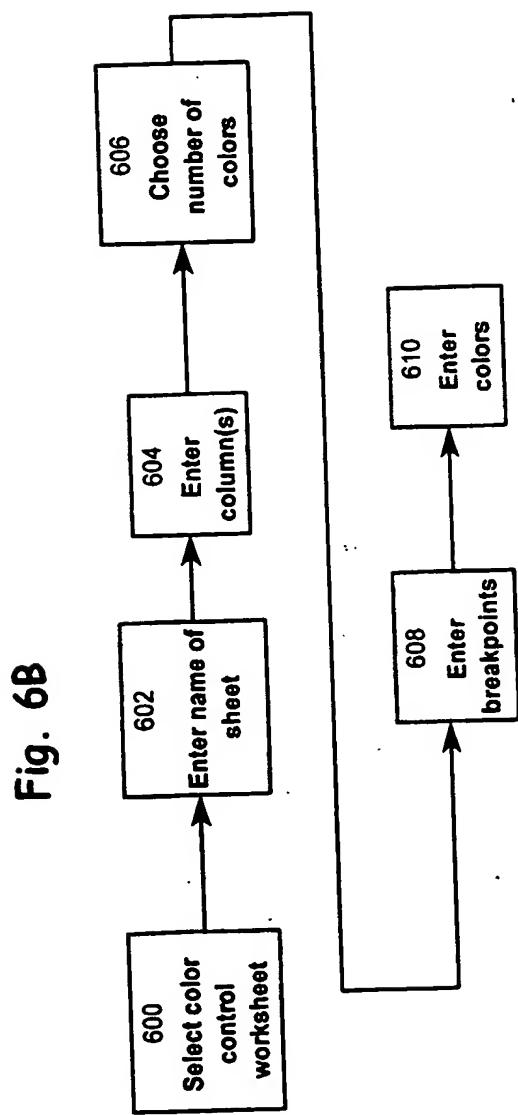


Fig. 6C

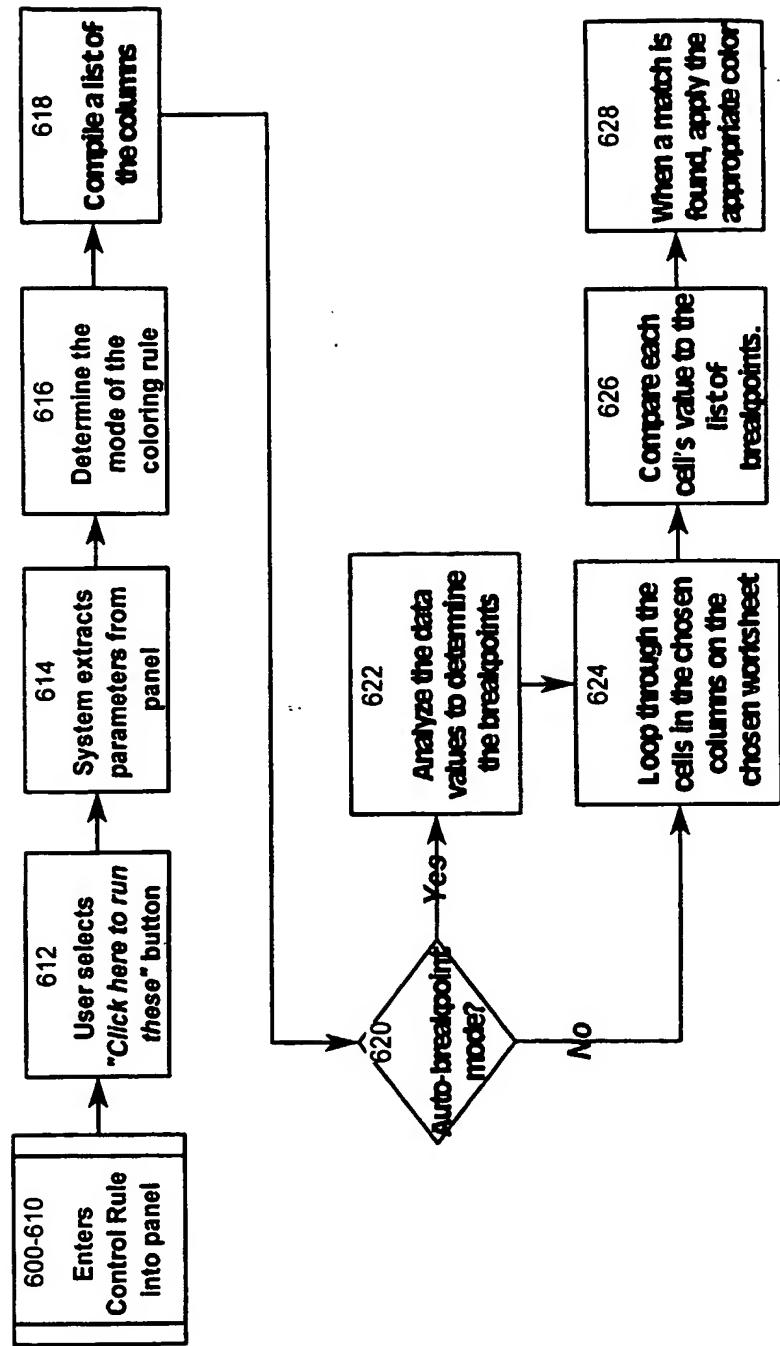


Fig. 7A

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	Cmpd01	N	23	30			3	22		6					
2	Cmpd02	N	42	55	41		57	28		15					
3	Cmpd03	G			11	83									
4	Cmpd04	N			30	70	26	24		29					
5	Cmpd05	N			92	89	60	21		22					
6	Cmpd06	D	8.36	6.5	37	71	41	13		3					
7	Cmpd07	D			7.8	100	79	48		43					
8	Cmpd08	D			2.6	65	28	28		38					
9	Cmpd09	D	0.119			68	41	22		15					
10	Cmpd10	N	0.233			61	42	24		5					
11	Cmpd11	N			50	77	63	25		3					
12	Cmpd12	H			0.24	47	25	24		37					
13	Cmpd13	H			0.194	30	81	69		40					
14	Cmpd14	H	0.28	0.41		39	23	4		12					
15	Cmpd15	H	0.369	0.168		89	46	46		36					
16	Cmpd16	I			0.87	30	101	82		38					
17	Cmpd17	K			0.223	30	81	64		47					
18	Cmpd18	I	5.27				79	54		22					
19	Cmpd19	I	0.134				71	71		23					
20	Cmpd20	F			0.317	101	109		108						
21	Cmpd21	K			2.21		87	70		31					
22	Cmpd22	B			0.15		94	77		36					
23	Cmpd23	B					96	61		36					
24	Cmpd24	B			0.27		110	91		43					
25	Cmpd25	B					105	104		75					
26	Cmpd26	B	0.665				93	71		41					
27	Cmpd27	B	0.111				97	79		43					
28	Cmpd28	E			0.13		95	93		52					
29	Cmpd29	J			0.46		68	62		12					
30	Cmpd30	J			4.5		41	48		9					
31	Cmpd31	N					52	6		21					
32	Cmpd32	J					62	30		15					
33	Cmpd33	N					112	29		75					
34	Cmpd34	J	1.227	2.4		30	105	62		62					
35	Cmpd35	I	0.63					44		28					
36		DEMO 1	DEMO 2	DEMO 3	DEMO 4	DEMO 5	DEMO 6	DEMO 7	DEMO 8	DEMO 9	DEMO 10	DEMO 11	DEMO 12	DEMO 13	DEMO 14

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
Cmpd	Series	Test1	Test2	Test3	HTS SPA % Inhib @ 3x10-6M	HTS SPA % Inhib @ 1x10-6M	HTS SPA % Inhib @ 3x10-7M	HTS SPA % Inhib @ 1x10-7M	HTS SPA % Inhib @ 3x10-7M	HTS SPA % Inhib @ 1x10-7M	HTS SPA % Inhib @ 3x10-7M	HTS SPA % Inhib @ 1x10-7M	HTS SPA % Inhib @ 3x10-7M	HTS SPA % Inhib @ 1x10-7M	
33	Cmpd32	J			62	20	75	15	33						
34	Cmpd33	N			112	23	62	62	21						
35	Cmpd34	J		30	105	44	26	18	3						
36	Cmpd35	J	0.63		93	58	20	28	30						
37	Cmpd36	J	0.23		90	60	61								
38	Cmpd37	N			97	73	23								
39	Cmpd38	L			65	56	40								
40	Cmpd39	F	0.56	0.61	68	46	28	13							
41	Cmpd40	G	0.388	0.35	102	87	92	87							
42	Cmpd41	G	0.038		38	26	18	12							
43	Cmpd42	F	0.38	0.9	78	60	40	25							
44	Cmpd43	H	0.747	0.13	111	110	104	82							
45	Cmpd44	G	0.05		25	21	16	21							
46	Cmpd45	F	0.6		100	104	97	80							
47	Cmpd46	M	0.117												
48	Cmpd47	F	0.117												
49	Cmpd48	F	0.27		100	102	93	75							
50	Cmpd49	M			71	43	31	15							
51	Cmpd50	F			112	11	112	75							
52	Cmpd51	G			78	70	44	27							
53	Cmpd52	H	0.33	0.117	69	24	23	14							
54	Cmpd53	A	0.117	0.18	1.6	1.08	1.02	0.63	27						
55	Cmpd54	A	0.33		91	78	76	33							
56	Cmpd55	G	0.112		61	64	60	26							
57	Cmpd56	C	0.39		93	80	55	37							
58	Cmpd57	A	0.117		101	72	42	29							
59	Cmpd58	C	0.22		92	69	55	49							

Fig. 7B

Fig. 8A

File Edit View Insert Format Tools Data Window Help

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
Cmpd	Series	Test1	Test2	Test3	HTS SPA Dose-Resp % Inhib @ 3x10-6M	HTS SPA Dose-Resp % Inhib @ 3x10-7M	HTS SPA Dose-Resp % Inhib @ 1x10-7M	HTS SPA Dose-Resp % Inhib @ 1x10-8M	HTS SPA Dose-Resp % Inhib @ 1x10-9M	HTS SPA Dose-Resp % Inhib @ 1x10-10M	HTS SPA Dose-Resp % Inhib @ 1x10-11M	HTS SPA Dose-Resp % Inhib @ 1x10-12M	HTS SPA Dose-Resp % Inhib @ 1x10-13M	HTS SPA Dose-Resp % Inhib @ 1x10-14M	
1	Cmpd01	N			23	30									
2	Cmpd02	N			42	65									
3	Cmpd03	G			11	83	67	28	5	15					
4	Cmpd04	N			20	70	25	24	29						
5	Cmpd05	N			32	88	60	21	22						
6	Cmpd06	D	8.86	5.5	27	71	12	13	3						
7	Cmpd07	D	0.119	0.119	25	79	49	49	49						
8	Cmpd08	D	0.119	0.119	26	65	26	26	38						
9	Cmpd09	D	0.119	0.119	68	4	22	22	15						
10	Cmpd10	N	0.233	0.233	61	42	24	24	5						
11	Cmpd11	N	0.117	0.117	74	77	63	63	25						
12	Cmpd12	H	0.24	0.24	81	69	40	40	37						
13	Cmpd13	H	0.194	0.194	30	33	23	23	3						
14	Cmpd14	H	0.026	0.026	41	48	46	46	36						
15	Cmpd15	H	0.369	0.369	0.118	82	38	38	18						
16	Cmpd16	I			0.07	30	81	64	37						
17	Cmpd17	K			0.0223	30	79	54	22						
18	Cmpd18	I	5.27			71	71	23	23						
19	Cmpd19	I	0.134			87	70	31	31						
20	Cmpd20	F		0.017	22	87	77	38	13						
21	Cmpd21	K				61	61	38	12						
22	Cmpd22	B		0.16			63	63	38						
23	Cmpd23	B			0.27		75	75	32						
24	Cmpd24	B					71	71	22						
25	Cmpd25	B					79	79	23						
26	Cmpd26	B	0.665				62	62	11						
27	Cmpd27	B	0.111				62	62	25						
28	Cmpd28	E		0.13			62	62	11						
29	Cmpd29	J		0.46			62	62	17						
30	Cmpd30	J		0.45			62	62	8						
31	Cmpd31	N					62	62	33						
32	Cmpd32	J					62	62	14						
33	Cmpd33	N					62	62	75						

DEM0 1 / DEM0 2 / DEM0 3 / distatinfo DEM0 / Appendix / Color Control / Cluster Control / NUM

Ready

Fig. 8B

File Edit View Insert Format Tools Date Window Help

A B C D E F G H I J K L M N O

Cmpd	Series	Test1			Test2			Test3			HTS SPA			HTS SPA			HTS SPA		
		Dose- Resp	% Inhib @ 3x10 ⁻⁶ M	Dose- Resp	% Inhib @ 3x10 ⁻⁶ M	Dose- Resp	% Inhib @ 3x10 ⁻⁶ M	Dose- Resp	% Inhib @ 3x10 ⁻⁷ M	Dose- Resp	% Inhib @ 3x10 ⁻⁷ M	Dose- Resp	% Inhib @ 1x10 ⁻⁷ M	Dose- Resp	% Inhib @ 1x10 ⁻⁷ M	Dose- Resp	% Inhib @ 1x10 ⁻⁷ M		
31	Cmpd34	J	22.55	24	30	44.55	47	57	28	62	62	21	3	14	30	30	30	30	
32	Cmpd35	J	0.63	0.23	0.11	0.41	0.11	0.09	0.21	0.31	0.31	0.21	0.11	0.09	0.11	0.11	0.11	0.11	
33	Cmpd36	N	12.3	10.1	10.1	10.1	10.1	10.1	10.1	10.1	10.1	10.1	10.1	10.1	10.1	10.1	10.1	10.1	
34	Cmpd37	N	1.23	1.23	1.23	1.23	1.23	1.23	1.23	1.23	1.23	1.23	1.23	1.23	1.23	1.23	1.23	1.23	
35	Cmpd38	L	0.56	0.41	0.35	0.35	0.35	0.35	0.35	0.35	0.35	0.35	0.35	0.35	0.35	0.35	0.35	0.35	
36	Cmpd39	F	0.38	0.11	0.09	0.09	0.09	0.09	0.09	0.09	0.09	0.09	0.09	0.09	0.09	0.09	0.09	0.09	
37	Cmpd40	G	0.38	0.11	0.09	0.09	0.09	0.09	0.09	0.09	0.09	0.09	0.09	0.09	0.09	0.09	0.09	0.09	
38	Cmpd41	G	0.38	0.09	0.09	0.09	0.09	0.09	0.09	0.09	0.09	0.09	0.09	0.09	0.09	0.09	0.09	0.09	
39	Cmpd42	F	0.38	0.09	0.09	0.09	0.09	0.09	0.09	0.09	0.09	0.09	0.09	0.09	0.09	0.09	0.09	0.09	
40	Cmpd43	H	4.16	0.13	0.13	0.13	0.13	0.13	0.13	0.13	0.13	0.13	0.13	0.13	0.13	0.13	0.13	0.13	
41	Cmpd44	G	0.029	0.029	0.029	0.029	0.029	0.029	0.029	0.029	0.029	0.029	0.029	0.029	0.029	0.029	0.029	0.029	
42	Cmpd45	F	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	
43	Cmpd46	M	0.114	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	
44	Cmpd47	F	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	
45	Cmpd48	M	0.015	0.027	0.027	0.027	0.027	0.027	0.027	0.027	0.027	0.027	0.027	0.027	0.027	0.027	0.027	0.027	
46	Cmpd49	F	0.015	0.027	0.027	0.027	0.027	0.027	0.027	0.027	0.027	0.027	0.027	0.027	0.027	0.027	0.027	0.027	
47	Cmpd50	G	0.017	0.027	0.027	0.027	0.027	0.027	0.027	0.027	0.027	0.027	0.027	0.027	0.027	0.027	0.027	0.027	
48	Cmpd51	H	0.33	0.117	0.49	0.69	0.69	0.69	0.69	0.69	0.69	0.69	0.69	0.69	0.69	0.69	0.69	0.69	
49	Cmpd52	A	0.061	0.10	0.16	0.16	0.16	0.16	0.16	0.16	0.16	0.16	0.16	0.16	0.16	0.16	0.16	0.16	
50	Cmpd53	A	0.33	0.33	0.33	0.33	0.33	0.33	0.33	0.33	0.33	0.33	0.33	0.33	0.33	0.33	0.33	0.33	
51	Cmpd54	A	0.39	0.12	0.12	0.12	0.12	0.12	0.12	0.12	0.12	0.12	0.12	0.12	0.12	0.12	0.12	0.12	
52	Cmpd55	G	0.39	0.22	0.22	0.22	0.22	0.22	0.22	0.22	0.22	0.22	0.22	0.22	0.22	0.22	0.22	0.22	
53	Cmpd56	C	0.039	0.039	0.039	0.039	0.039	0.039	0.039	0.039	0.039	0.039	0.039	0.039	0.039	0.039	0.039	0.039	
54	Cmpd57	A	0.22	0.22	0.22	0.22	0.22	0.22	0.22	0.22	0.22	0.22	0.22	0.22	0.22	0.22	0.22	0.22	
55	Cmpd58	C	0.022	0.022	0.022	0.022	0.022	0.022	0.022	0.022	0.022	0.022	0.022	0.022	0.022	0.022	0.022	0.022	
56																			
57																			
58																			
59																			
60																			
61																			
62																			
63																			
64																			
65																			
66																			
67																			
68																			
69																			
70																			
71																			

Ready

DEMO 1 / DEMO 2 / DEMO 3 / Apcard Control / Color Control / Cluster Control / NUM

DEMO 1

Fig. 9B

Fig. 10A

Fig. 10B

Microsoft Excel - demo~PANDORA~v23.1.xls [Read-Only]

File	Edit	View	User	Format	Tools	Data	Window	Help
F2 A	F2 B	F2 C	A	D	32			
F2 B	F2 F	F2 C	B					
F2 C	F2 F	F2 C	C					
F2 D	F2 F	F2 D	D					
F3 A	F3 F	F3 A	A					
F3 B	F3 F	F3 B	B		39			
F3 C	F3 F	F3 C	C		33			
F3 D	F3 F	F3 D	D		42			
F4 A	F4 F	F4 A	A		41			
F4 B	F4 F	F4 B	B		3			
F4 C	F4 F	F4 C	C		19			
F4 D	F4 F	F4 D	D		24			
F5 A	F5 F	F5 A	A		1			
F5 B	F5 F	F5 B	B		21			
F5 C	F5 F	F5 C	C		23			
F5 D	F5 F	F5 D	D		22			
F5 E	F5 F	F5 E	E		19			
F5 F	F5 F	F5 F	F		35			
F6 A	F6 F	F6 A	A		19			
F6 B	F6 F	F6 B	B		35			
F6 C	F6 F	F6 C	C		26			
F6 D	F6 F	F6 D	D		11			
F7 A	F7 F	F7 A	A		4			
F7 B	F7 F	F7 B	B		19			
F7 C	F7 F	F7 C	C		9			
F7 D	F7 F	F7 D	D		15			
F8 E	F8 F	F8 E	E		16			
F8 F	F8 F	F8 F	F		16			
F9 G	F9 F	F9 G	G		16			
F9 H	F9 F	F9 H	H		16			
F9 I	F9 F	F9 I	I		16			
F9 J	F9 F	F9 J	J		16			
F9 K	F9 F	F9 K	K		16			
F9 L	F9 F	F9 L	L		16			
F9 M	F9 F	F9 M	M		16			
F9 N	F9 F	F9 N	N		16			
F9 O	F9 F	F9 O	O		16			
F9 P	F9 F	F9 P	P		16			
F9 Q	F9 F	F9 Q	Q		16			
F9 R	F9 F	F9 R	R		16			
F9 S	F9 F	F9 S	S		16			
F9 T	F9 F	F9 T	T		16			
F9 U	F9 F	F9 U	U		16			
F9 V	F9 F	F9 V	V		16			
F9 W	F9 F	F9 W	W		16			
F9 X	F9 F	F9 X	X		16			
F9 Y	F9 F	F9 Y	Y		16			
F9 Z	F9 F	F9 Z	Z		16			

Fig. 11A 314

Sheet 1		Sheet 2		Sheet 3	
Columns	F:I	Columns	F:I	Columns	F:I
Name:	acids	Name:	mines	Name:	metals
Sheet #	DEMO 2	Sheet #	DEMO 3	Sheet #	DEMO 4
Cluster Col	B	Cluster Col	C	Cluster Col	D
Color	red	Color	red	Color	red
Score	0	Score	0	Score	0
Color	orange	Color	orange	Color	orange
Score	1	Score	1	Score	1
Color	yellow	Color	yellow	Color	yellow
Score	2	Score	2	Score	2
Color	blue	Color	blue	Color	blue
Score	3	Score	3	Score	3
Column (b)	Rel. Weight	Column (b)	Rel. Weight	Column (b)	Rel. Weight
D	1	D	1	D	1
Column (b)	Rel. Weight	Column (b)	Rel. Weight	Column (b)	Rel. Weight
E	1	E	1	E	1
Column (b)	Rel. Weight	Column (b)	Rel. Weight	Column (b)	Rel. Weight
F	1	F	1	F	1
Column (b)	Rel. Weight	Column (b)	Rel. Weight	Column (b)	Rel. Weight
G	1	G	1	G	1
Column (b)	Rel. Weight	Column (b)	Rel. Weight	Column (b)	Rel. Weight
H	1	H	1	H	1
Column (b)	Rel. Weight	Column (b)	Rel. Weight	Column (b)	Rel. Weight
I	1	I	1	I	1

Fig. 11B

1100

1102

1104

1106

1108

1110

1112

1114

1116

Name:	acids
Sheet #	DEMO2
Cluster Col	B

Color	Score
red	0
yellow	2

Column #	Ref. Weight
0	1

Fig. 11C

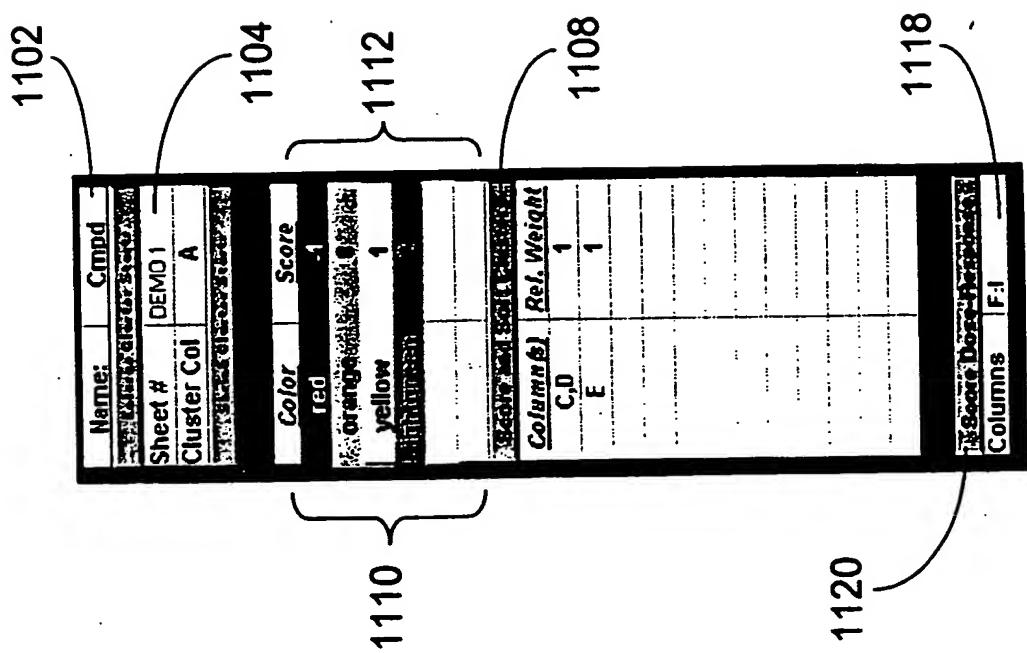


Fig. 11D

B1		acid	
		amine	
		group	
		acid	

E1		acid	
E1_A	A		
E1_B	E1	B	
E1_C	E1	C	40
E1_D	E1	D	
E2		acid	
E2_A	A	7	
E2_B	E2	B	
E2_C	E2	C	40
E2_D	E2	D	
E3		acid	
E3_A	A	1	
E3_B	E3	B	2
E3_C	E3	C	20
E3_D	E3	D	14
E4		acid	
E4_A	A	2	
E4_B	E4	B	15
E4_C	E4	C	27
E4_D	E4	D	12
E5		acid	
E5_A	A	3	
E5_B	E5	B	5
E5_C	E5	C	28
E5_D	E5	D	10
E6		acid	
E6_A	A	5	

Fig. 13A

Fig. 13B

318

Fig. 13c

306

Fig. 13D

A screenshot of the Microsoft Word ribbon interface. The ribbon tabs visible are File, Edit, View, Insert, Format, Tools, Data, Start, Window, Transfer, PANDORA, Help, and Arial. The Arial font is selected in the font dropdown. The status bar at the bottom shows the text 'Arial' and the number '10'.

Fig. 14A

314 1402

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
Cmpd	Series	Test1	Test2	Test3	Dose-Resp	HTS SPA	HTS SPA	HTS SPA	Dose-Resp	Dose-Resp	Dose-Resp	HTS SPA			
					% Inhib @ 3x10-6M	% Inhib @ 1x10-6M	% Inhib @ 3x10-7M	% Inhib @ 1x10-7M	% Inhib @ 3x10-6M	% Inhib @ 1x10-6M	% Inhib @ 3x10-7M				
1	Cmpd38	L				90	90	90							
2	Cmpd44	C				78	78	78							
3	Cmpd49	M				71	71	71							
4	Cmpd53	A			16	16	16								
5	Cmpd53	A			26	26	26								
6	Cmpd08	D				65	65	65							
7	Cmpd41	G				66	66	66							
8	Cmpd52	H			49	49	49								
9	Cmpd07	D			69	69	69								
10	Cmpd14	H			24	24	24								
11	Cmpd15	H			41	41	41								
12	Cmpd26	B			0.169	0.169	0.169								
13	Cmpd39	F			0.41	0.41	0.41								
14	Cmpd42	F			0.38	0.38	0.38								
15	Cmpd47	F			0.9	0.9	0.9								
16	Cmpd50	F													
17	Cmpd51	G													
18	Cmpd57	A													
19	Cmpd09	D			0.119	0.119	0.119								
20	Cmpd10	N			0.233	0.233	0.233								
21	Cmpd12	H			0.10	0.10	0.10								
22	Cmpd19	I			0.134	0.134	0.134								
23	Cmpd20	F			0.317	0.317	0.317								
24	Cmpd22	B			0.16	0.16	0.16								
25	Cmpd24	B			0.27	0.27	0.27								
26	Cmpd26	B			0.665	0.665	0.665								
27	Cmpd27	B			0.111	0.111	0.111								
28	Cmpd28	E			0.13	0.13	0.13								
29	Cmpd29	J			0.46	0.46	0.46								
30	Cmpd35	J			0.63	0.63	0.63								
31	Cmpd36	J			0.23	0.23	0.23								
32	Cmpd40	G			0.368	0.368	0.368								
33	Cmpd43	H			0.113	0.113	0.113								
34	Cmpd45	F			0.5	0.5	0.5								

Microsoft Excel

OK

Elapsed time 5 seconds

Two sheets have been added to the current workbook:
'DEMO 1 SCORES by Cmpd'
'DEMO 1 SORTED by Cmpd'

Summarize data

Two sheets have been added to the current workbook:
'DEMO 1 SCORES by Cmpd'
'DEMO 1 SORTED by Cmpd'

OK

Elapsed time 5 seconds

Two sheets have been added to the current workbook:
'DEMO 1 SCORES by Cmpd un=ze'
'DEMO 2 / DEMO 3 / DEMO 4'

OK

Elapsed time 5 seconds

Two sheets have been added to the current workbook:
'DEMO 1 SCORES by Cmpd un=ze'
'DEMO 2 / DEMO 3 / DEMO 4'

OK

Elapsed time 5 seconds

Two sheets have been added to the current workbook:
'DEMO 1 SCORES by Cmpd un=ze'
'DEMO 2 / DEMO 3 / DEMO 4'

OK

Elapsed time 5 seconds

Two sheets have been added to the current workbook:
'DEMO 1 SCORES by Cmpd un=ze'
'DEMO 2 / DEMO 3 / DEMO 4'

OK

Elapsed time 5 seconds

Two sheets have been added to the current workbook:
'DEMO 1 SCORES by Cmpd un=ze'
'DEMO 2 / DEMO 3 / DEMO 4'

OK

Elapsed time 5 seconds

Two sheets have been added to the current workbook:
'DEMO 1 SCORES by Cmpd un=ze'
'DEMO 2 / DEMO 3 / DEMO 4'

OK

Elapsed time 5 seconds

Two sheets have been added to the current workbook:
'DEMO 1 SCORES by Cmpd un=ze'
'DEMO 2 / DEMO 3 / DEMO 4'

OK

Fig. 14B

Fig. 14C

File Edit View Insert Format Tools Data Window Help

A B C D E F G H I J K L M N O P

cluster # of cmpds (undesired score max 100 (max 100 <-- Scoring of sheet "DEMO 1" using parameter set "Cmpd"

label = zero)

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37
Cmpd38	1	67	100																																	
Cmpd44	1	67	100																																	
Cmpd49	1	67	100																																	
Cmpd55	1	67	67																																	
Cmpd68	1	50	75																																	
Cmpd41	1	50	50																																	
Cmpd52	1	50	60																																	
Cmpd07	1	33	33																																	
Cmpd11	1	33	33																																	
Cmpd15	1	33	33																																	
Cmpd25	1	33	33																																	
Cmpd39	1	33	33																																	
Cmpd42	1	33	33																																	
Cmpd47	1	33	33																																	
Cmpd50	1	33	33																																	
Cmpd51	1	33	33																																	
Cmpd57	1	33	33																																	
Cmpd09	1	17	17																																	
Cmpd10	1	17	17																																	
Cmpd12	1	17	17																																	
Cmpd19	1	17	20																																	
Cmpd20	1	17	20																																	
Cmpd22	1	17	21																																	
Cmpd24	1	17	21																																	
Cmpd26	1	17	21																																	
Cmpd27	1	17	21																																	
Cmpd28	1	17	21																																	
Cmpd29	1	17	21																																	
Cmpd35	1	17	21																																	
Cmpd36	1	17	21																																	
Cmpd40	1	17	21																																	
Cmpd43	1	17	21																																	
Cmpd45	1	17	21																																	
Cmpd48	1	17	21																																	
Cmpd54	1	17	21																																	
Cmpd56	1	17	21																																	
Cmpd57	1	17	21																																	

DEMO 1 SCORES by Cmpd # DEMO 1 SORTED by Cmpd # DEMO 1 FILTERED by Cmpd # DEMO 1

Ready

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z
30	Cmp035	1	17	20	20	20	20	20	20	20	20	20	20	20	20	20	20	20	20	20	20	20	20	20	20	20
31	Cmp036	1	17	20	20	20	20	20	20	20	20	20	20	20	20	20	20	20	20	20	20	20	20	20	20	20
32	Cmp040	1	17	20	20	20	20	20	20	20	20	20	20	20	20	20	20	20	20	20	20	20	20	20	20	20
33	Cmp043	1	17	20	20	20	20	20	20	20	20	20	20	20	20	20	20	20	20	20	20	20	20	20	20	20
34	Cmp045	1	17	20	20	20	20	20	20	20	20	20	20	20	20	20	20	20	20	20	20	20	20	20	20	20
35	Cmp049	1	17	20	20	20	20	20	20	20	20	20	20	20	20	20	20	20	20	20	20	20	20	20	20	20
36	Cmp054	1	17	20	20	20	20	20	20	20	20	20	20	20	20	20	20	20	20	20	20	20	20	20	20	20
37	Cmp056	1	17	20	20	20	20	20	20	20	20	20	20	20	20	20	20	20	20	20	20	20	20	20	20	20
38	Cmp058	1	17	20	20	20	20	20	20	20	20	20	20	20	20	20	20	20	20	20	20	20	20	20	20	20
39	Cmp059	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
40	Cmp051	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
41	Cmp013	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
42	Cmp016	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
43	Cmp017	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
44	Cmp021	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
45	Cmp023	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
46	Cmp031	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
47	Cmp032	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
48	Cmp033	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
49	Cmp046	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
50	Cmp002	32	32	32	32	32	32	32	32	32	32	32	32	32	32	32	32	32	32	32	32	32	32	32	32	32
51	Cmp003	32	32	32	32	32	32	32	32	32	32	32	32	32	32	32	32	32	32	32	32	32	32	32	32	32
52	Cmp004	32	32	32	32	32	32	32	32	32	32	32	32	32	32	32	32	32	32	32	32	32	32	32	32	32
53	Cmp006	32	32	32	32	32	32	32	32	32	32	32	32	32	32	32	32	32	32	32	32	32	32	32	32	32
54	Cmp018	32	32	32	32	32	32	32	32	32	32	32	32	32	32	32	32	32	32	32	32	32	32	32	32	32
55	Cmp030	32	32	32	32	32	32	32	32	32	32	32	32	32	32	32	32	32	32	32	32	32	32	32	32	32
56	Cmp037	32	32	32	32	32	32	32	32	32	32	32	32	32	32	32	32	32	32	32	32	32	32	32	32	32
57	Cmp055	67	67	67	67	67	67	67	67	67	67	67	67	67	67	67	67	67	67	67	67	67	67	67	67	67
58	Cmp001	67	67	67	67	67	67	67	67	67	67	67	67	67	67	67	67	67	67	67	67	67	67	67	67	67
59	Cmp034	67	67	67	67	67	67	67	67	67	67	67	67	67	67	67	67	67	67	67	67	67	67	67	67	67
60		61	61	61	61	61	61	61	61	61	61	61	61	61	61	61	61	61	61	61	61	61	61	61	61	61
61		62	62	62	62	62	62	62	62	62	62	62	62	62	62	62	62	62	62	62	62	62	62	62	62	62
62		63	63	63	63	63	63	63	63	63	63	63	63	63	63	63	63	63	63	63	63	63	63	63	63	63
63		64	64	64	64	64	64	64	64	64	64	64	64	64	64	64	64	64	64	64	64	64	64	64	64	64
64		65	65	65	65	65	65	65	65	65	65	65	65	65	65	65	65	65	65	65	65	65	65	65	65	65
65		66	66	66	66	66	66	66	66	66	66	66	66	66	66	66	66	66	66	66	66	66	66	66	66	66
66		67	67	67	67	67	67	67	67	67	67	67	67	67	67	67	67	67	67	67	67	67	67	67	67	67
67		68	68	68	68	68	68	68	68	68	68	68	68	68	68	68	68	68	68	68	68	68	68	68	68	68
68		69	69	69	69	69	69	69	69	69	69	69	69	69	69	69	69	69	69	69	69	69	69	69	69	69
69		70	70	70	70	70	70	70	70	70	70	70	70	70	70	70	70	70	70	70	70	70	70	70	70	70
70		71	71	71	71	71	71	71	71	71	71	71	71	71	71	71	71	71	71	71	71	71	71	71	71	71

Fig. 14D

DEMO 1 SCORES by Cmpd / DEMO 1 SORTED by Cmpd / DEMO 1 Unsort

Fig. 15A

	A	B	C	D	E	F	G	H	I	J	K
1	mixture	amine group	acid	activity							
2	E1 A	E1	A	26							
3	E1 B	E1	B	40							
4	E1 C	E1	C	32							
5	E1 D	E1	D	42							
6	F2 A	F2	A	35							
7	F2 B	F2	B	26							
8	F2 C	F2	C	42							
9	F2 D	F2	D	34							
10	E7 A	E7	A	35							
11	E7 B	E7	B	41							
12	E7 C	E7	C	33							
13	E7 D	E7	D	42							
14	F1 A	F1	A	42							
15	F1 B	F1	B	26							
16	F1 C	F1	C	35							
17	F1 D	F1	D	41							
18	F6 A	F6	A	38							
19	F6 B	F6	B	33							
20	F6 C	F6	C	42							
21	F6 D	F6	D	35							
22	F3 A	F3	A	38							
23	F3 B	F3	B	33							
24	F3 C	F3	C	42							
25	F3 D	F3	D	35							
26	E2 A	E2	A	7							
27	E2 B	E2	B	40							
28	E2 C	E2	C	35							
29	E2 D	E2	D	42							
30	E3 A	E3	A	38							
31	E3 B	E3	B	33							
32	E3 C	E3	C	42							
33	E3 D	E3	D	35							
34	E4 A	E4	A	38							
35	E4 B	E4	B	33							
36	E4 C	E4	C	42							
37	E4 D	E4	D	35							
38	E5 A	E5	A	38							
39	E5 B	E5	B	33							
40	E5 C	E5	C	42							
41	E5 D	E5	D	35							
42	E6 A	E6	A	38							
43	E6 B	E6	B	33							
44	E6 C	E6	C	42							
45	E6 D	E6	D	35							

Fig. 15B

	A	B	C	D	E
Cmpd	HTS SPA Dose-Resp % Inhib @ 3x10 ⁻⁶ M	HTS SPA Dose-Resp % Inhib @ 1x10 ⁻⁶ M	HTS SPA Dose-Resp % Inhib @ 3x10 ⁻⁷ M	HTS SPA Dose-Resp % Inhib @ 1x10 ⁻⁷ M	HTS SPA Dose-Resp % Inhib @ 1x10 ⁻⁷ M
1					
2	Cmpd01	41	22	16	19
3	Cmpd02	57	22	5	5
4	Cmpd03	25	28	15	15
5	Cmpd04	60	24	29	29
6	Cmpd05	41	21	22	22
7	Cmpd06	71	13	3	3
8	Cmpd07	71	48	43	43
9	Cmpd08	65	28	38	38
10	Cmpd09	58	41	22	15
11	Cmpd10	61	42	24	5
12	Cmpd11	50	77	63	25
13	Cmpd12	47	25	24	3
14	Cmpd13	61	59	40	37
15	Cmpd14	39	23	4	12
16	Cmpd15	59	46	46	36
17	Cmpd16	101	65	38	18
18	Cmpd17	79	64	47	24
19	Cmpd18	71	54	22	32
20	Cmpd19	71	71	23	12
21	Cmpd20	101	71	71	12

Figure 16A

Click here to run these

sheet	DEMO 1
column(s)	B:E
# of colors	3
break 1	33
break 2	67
break 3	

color 1 red

color 2 yellow

color 3 black

Re-scale all?

Figure 16B

Fig. 16C

Cmpd	A	B	C	D	E	F	G	H
	HTS SPA Dose-Resp % Inhib @ 3x10-6M 3.00e-06	HTS SPA Dose-Resp % Inhib @ 1x10-6M 1.00e-06	HTS SPA Dose-Resp % Inhib @ 3x10-7M 3.00e-07	HTS SPA Dose-Resp % Inhib @ 1x10-7M 1.00e-07	D-R -iveness score (0 to 100) by Cmpd	D-R activity score (0 to 100) by Cmpd	D-R composite score (0 to 100) by Cmpd	D-R composite score (0 to 100) by Cmpd
1					75	100	100	100
2	Cmpd20				83	77	79	75
3	Cmpd07				83	70	75	75
4	Cmpd13				83	70	75	75
5	Cmpd15				83	70	70	70
6	Cmpd16				92	50	67	67
7	Cmpd17				92	43	57	57
8	Cmpd03				92	23	57	57
9	Cmpd05				92	23	57	57
10	Cmpd06				92	23	57	57
11	Cmpd09				92	23	57	57
12	Cmpd11	50			67	47	57	57
13	Cmpd18	70	54		92	23	57	57
14	Cmpd19	71			92	30	56	56
15	Cmpd10	61	42		5	20	51	46
16	Cmpd01	70	22		19	83	10	46
17	Cmpd04	70	25		29	83	7	45
18	Cmpd02	41	3		5	83	33	45
19	Cmpd08	65	28		38	3	7	45
20	Cmpd12	47	25		24	83	7	45
21	Cmpd14	39	23		4	12	83	7

	A	B	C	D	E	F	G	H
Cmpd	HTS SPA Dose-Resp % Inhib @ 3x10-6M 3.00e-06	HTS SPA Dose-Resp % Inhib @ 1x10-6M 1.00e-06	HTS SPA Dose-Resp % Inhib @ 3x10-7M 3.00e-07	HTS SPA Dose-Resp % Inhib @ 1x10-7M 1.00e-07	D-R -iveness score (0 to 100) by Cmpd	D-R activity score (0 to 100) by Cmpd	D-R compositi score (0 to 100) by Cmpd	D-R
1					75	100	100	100
22	marker_7.5				50	83	87	86
23	marker_7.0				49	92	50	70
24	marker_6.5				24	92	23	57
25	marker_6.0				9	92	7	45
26	marker_5.5				3	83	0	38
27	marker_5.0				1	75		

Figure 16D

A	B	C	D	E	F	G	H	I	J
	HTS SPA Dose-Resp % Inhib @ 3x10-6M 3.00e-06	HTS SPA Dose-Resp % Inhib @ 1x10-6M 1.00e-06	HTS SPA Dose-Resp % Inhib @ 3x10-7M 3.00e-07	HTS SPA Dose-Resp % Inhib @ 1x10-7M 1.00e-07	D-R activity score (0 to 100) by Cmpd	D-R activity score (0 to 100) by Cmpd	D-R composite score (0 to 100) by Cmpd	interp -log IC50 μM by Cmpd	est IC50 μM by Cmpd
1	Cmpd20 marker_7.5				75	100	100	<0.1	
2					75	100	100		
3					63	87	86		
4	marker_7.0				48	43	79	6.78	0.17
5	Cmpd07	59	40	37	63	77	75	6.66	0.22
6	Cmpd13	46	46	36	83	70	75	6.65	0.22
7	Cmpd15				18	92	50	6.50	0.32
8	Cmpd16				24	92	50	70	
9	marker_6.5	49	49	47	43	43	67	6.38	0.41
10	Cmpd17	64	28	15	92	23	57	6.00	1
11	Cmpd03	57	21	22	92	23	57	6.00	1
12	Cmpd05	60	13	3	92	23	57	6.00	1
13	Cmpd06	41	22	15	92	23	57	6.00	1
14	Cmpd09	41	63	25	67	47	57	6.00	1
15	Cmpd11	50			32	92	23	57	6.00
16	Cmpd18	54	22	9	92	23	57	5.75	1.8
17	marker_6.0	50	23	12	83	30	56	5.96	1.1
18	Cmpd19				5	19	20	51	5.54
19	Cmpd00	61	42	24	16	24	29	46	5.54
20	Cmpd01		22	5	63	10	10	46	2.9
21	Cmpd04		25	22	83	10	7	45	2.9
22	Cmpd02	41	3	22	5	83	7	45	>3
23	Cmpd08	65	28	28	58	33	45	7	>3
24	Cmpd12	47	25	24	3	63	7	45	>3
25	Cmpd14	39	23	4	12	3	7	45	>3
26	marker_5.5	49	24	9	63	7	7	0	
27	marker_5.0	23	9	3	1	75	0	38	

Figure 16E

Fig. 16F

Table 2. The complete data set for 3 points and 3 colors, in systematic order.

compound	percent inhibition			data group number			data group color		
	highest conc	lowest conc	--->	highest conc	lowest conc	--->	highest conc	lowest conc	--->
cmpd 01	2	29	22	1	1	1	1	2	1
cmpd 02	15	10	38	1	1	1	1	3	2
cmpd 03	31	26	79	1	1	2	2	2	2
cmpd 04	21	46	25	1	1	2	2	3	3
cmpd 05	30	53	61	1	1	2	3	1	1
cmpd 06	17	37	70	1	1	3	1	2	3
cmpd 07	26	90	17	1	1	3	2	3	3
cmpd 08	10	90	38	1	1	3	2	3	2
cmpd 09	32	72	94	1	1	3	1	1	1
cmpd 10	34	17	15	2	2	1	2	1	2
cmpd 11	51	8	42	2	2	1	3	1	3
cmpd 12	56	3	69	2	2	1	1	2	1
cmpd 13	33	39	14	2	2	2	2	2	2
cmpd 14	53	52	55	2	2	2	3	1	1
cmpd 15	51	52	87	2	2	3	1	2	3
cmpd 16	65	82	6	2	3	2	3	2	3
cmpd 17	43	71	41	2	3	3	3	2	3
cmpd 18	65	99	77	2	3	1	1	2	3
cmpd 19	67	11	9	3	1	1	2	1	1
cmpd 20	87	5	66	3	3	1	3	1	3
cmpd 21	77	8	70	3	3	2	2	2	2
cmpd 22	78	36	0	3	3	2	3	2	3
cmpd 23	85	40	63	3	3	2	3	3	2
cmpd 24	83	57	88	3	3	2	3	1	2
cmpd 25	73	88	15	3	3	3	3	3	3
cmpd 26	69	85	35	3	3	3	3	3	3
cmpd 27	79	68	91	3	3	3	3	3	3

Fig. 166

Table 3. The complete data set for 3 points and 3 colors, sorted by decreasing dose-responsiveness

compound	highest conc	... >	lowest conc	step scoring	scaled responsiveness 0-100
cmpd 22				+1+1	2 100
cmpd 10				+1+0	1 88
cmpd 13				0+1	1 88
cmpd 19				+1+0	1 88
cmpd 23				+1+0	1 88
cmpd 25				0+1	1 88
cmpd 26				0+1	1 88
cmpd 01				0+0	0 75
cmpd 14				0+0	0 75
cmpd 27				0+0	0 75
cmpd 04				-3+1	.2 50
cmpd 07				-3+1	.2 50
cmpd 08				-3+1	.2 50
cmpd 11				+1-3	.2 50
cmpd 12				+1-3	.2 50
cmpd 16				-3+1	.2 50
cmpd 17				-3+1	.2 50
cmpd 20				+1-3	.2 50
cmpd 21				+1-3	.2 50
cmpd 24				+1-3	.2 50
cmpd 02				0-3	.3 38
cmpd 03				0-3	.3 38
cmpd 05				-3+0	.3 38
cmpd 09				-3+0	.3 38
cmpd 15				0-3	.3 38
cmpd 18				-3+0	.3 38
cmpd 06				-3-3	.6 0

Fig. 16H

Table 4. The complete set of data for 3 points and 3 colors, sorted by decreasing overall activity.

compound	highest conc	→	lowest conc	data group number	data group color	lowest conc	→	highest conc	activity scoring	unscaled activity points	scaled activity 0-100
cmpd 27	3	3	3						1(3)*2(3)+3(3)	18	100
cmpd 18	2	3	3						1(2)*2(3)+3(3)	17	92
cmpd 24	3	2	3						1(3)*2(2)+3(3)	16	83
cmpd 09	1	3	3						1(1)*2(3)+3(3)	16	83
cmpd 26	3	3	2						1(3)*2(3)+3(2)	15	75
cmpd 25	2	2	3						1(2)*2(2)+3(3)	15	75
cmpd 15	2	3	2						1(2)*2(3)+3(2)	14	67
cmpd 17	2	3	2						1(3)*2(1)+3(3)	14	67
cmpd 21	3	1	3						1(1)*2(2)+3(3)	14	67
cmpd 06	1	2	3						1(1)*2(2)+3(3)	14	67
cmpd 23	3	2	2						1(3)*2(2)+3(2)	13	58
cmpd 08	1	3	2						1(1)*2(3)+3(2)	13	58
cmpd 12	2	1	3						1(2)*2(1)+3(3)	13	58
cmpd 25	3	3	1						1(3)*2(3)+3(1)	12	50
cmpd 14	2	2	2						1(2)*2(2)+3(2)	12	50
cmpd 03	1	1	3						1(1)*2(1)+3(3)	12	50
cmpd 16	2	3	1						1(2)*2(3)+3(1)	11	42
cmpd 20	3	1	2						1(3)*2(1)+3(2)	11	42
cmpd 05	1	2	2						1(1)*2(2)+3(2)	11	42
cmpd 22	3	2	1						1(3)*2(2)+3(1)	10	33
cmpd 07	1	3	1						1(1)*2(3)+3(1)	10	33
cmpd 11	2	1	2						1(2)*2(1)+3(2)	10	33
cmpd 13	2	2	1						1(2)*2(2)+3(1)	9	25
cmpd 02	1	1	2						1(1)*2(1)+3(2)	9	25
cmpd 19	3	1	1						1(3)*2(1)+3(1)	8	17
cmpd 04	1	2	1						1(1)*2(2)+3(1)	8	17
cmpd 10	2	1	1						1(2)*2(1)+3(1)	7	8
cmpd 01	1								1(1)*2(1)+3(1)	6	0

Fig. 16I

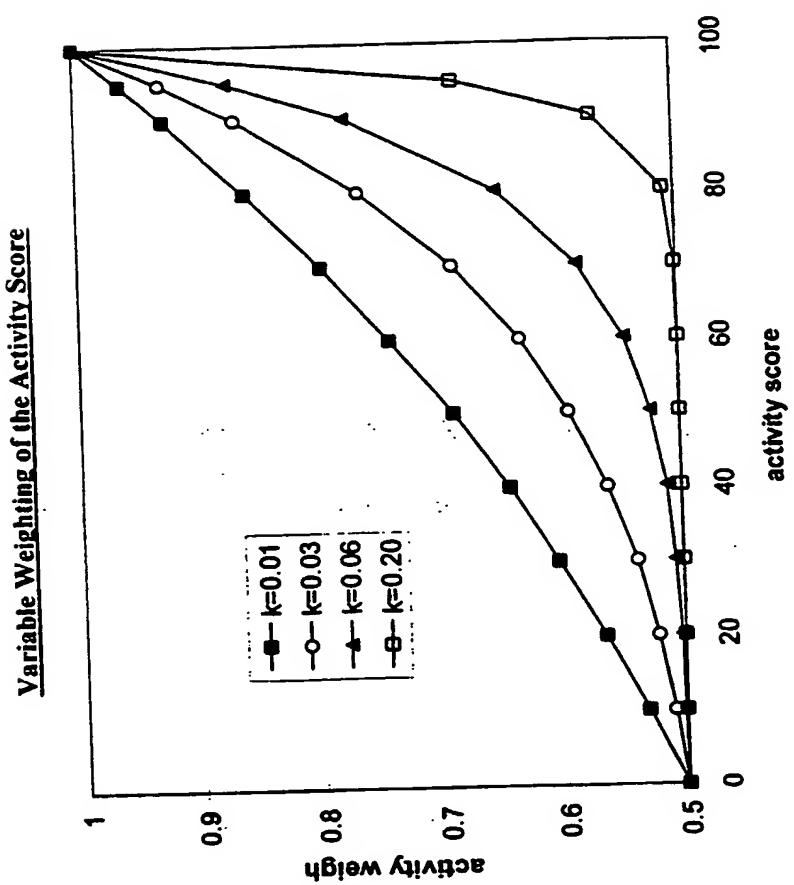


Fig. 16J

Table 5. The complete set of data for 3 points and 3 colors, sorted by decreasing composite score.

compound	highest conc	→	lowest conc	scaled responsive-ness 0-100	scaled activity 0-100	composite 0-100
cmpd 27				75	100	100
cmpd 18				38	92	82
cmpd 26				88	75	80
cmpd 24				50	83	72
cmpd 23				88	58	72
cmpd 09				38	83	69
cmpd 25				88	50	68
cmpd 22				100	33	66
cmpd 14				75	50	62
cmpd 15				38	75	61
cmpd 17				50	67	60
cmpd 21				50	67	60
cmpd 13				88	25	56
cmpd 08				50	58	54
cmpd 12				50	58	54
cmpd 19				88	17	52
cmpd 10				88	8	48
cmpd 16				50	42	46
cmpd 20				50	42	46
cmpd 03				38	50	44
cmpd 07				50	33	41
cmpd 11				50	33	41
cmpd 05				38	42	40
cmpd 06				0	67	38
cmpd 01				75	0	38
cmpd 04				50	17	33
cmpd 02				38	25	

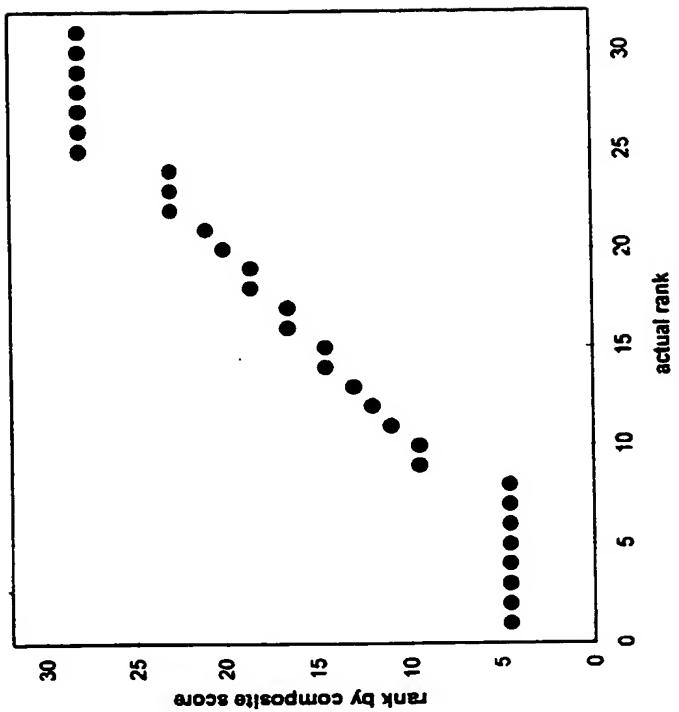
Fig. 16K**Quality of Ranking when Noise = 10 Inhibition Percentage Points**

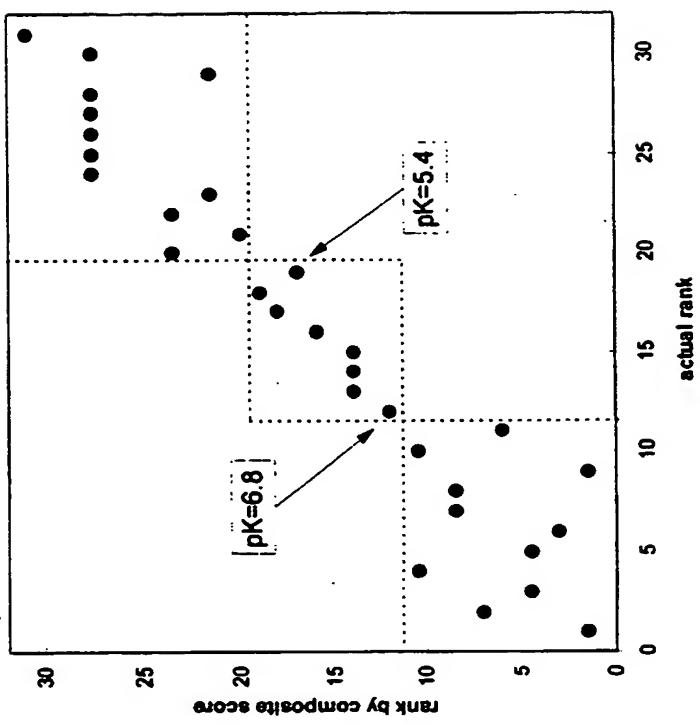
Fig. 16L**Quality of Ranking when Noise = 30 Inhibition Percentage Points**

Fig. 16M

Table 6. Quantitative Estimation of Potencies by Calibration Marker Compounds

compound	% Inhib @ 3.00e-06	% Inhib @ 1.00e-06	% Inhib @ 3.00e-07	% Inhib @ 1.00e-07	D-R composite score	Interp est log IC50 μ M
M363875	100	102	83	75	79	<0.1
marker_7.5	99	97	90	76	79	
M221211	110	91	69	39	76	7.00 0.10
M371565	108	102	63	27	76	7.00 0.10
marker_7.0	97	91	75	50	76	
M345077	102	67	92	87	75	6.97 0.11
M371756	91	78	76	33	73	6.90 0.13
M143629	100	79	48	43	69	6.77 0.17
M371890	101	72	42	29	69	6.77 0.17
M371891	92	69	55	49	69	6.77 0.17
M309032	105	62	62	21	67	6.70 0.20
M198289	101	82	38	18	66	6.67 0.22
M224602	97	79	43	23	66	6.67 0.22
M318571	93	63	42	14	66	6.67 0.22
M273373	95	93	52	25	65	6.63 0.23
M371336	78	70	44	27	62	6.53 0.29
M371825	61	64	60	26	62	6.53 0.29
M181250	99	46	46	36	61	6.50 0.32
M336331	87	73	29	15	61	6.50 0.32
marker_6.5	90	76	49	24	61	
M143630	65	28	28	38	60	6.44 0.36

Fig. 16N

Quality of Estimation when Noise = 10 Inhibition Percentage Points

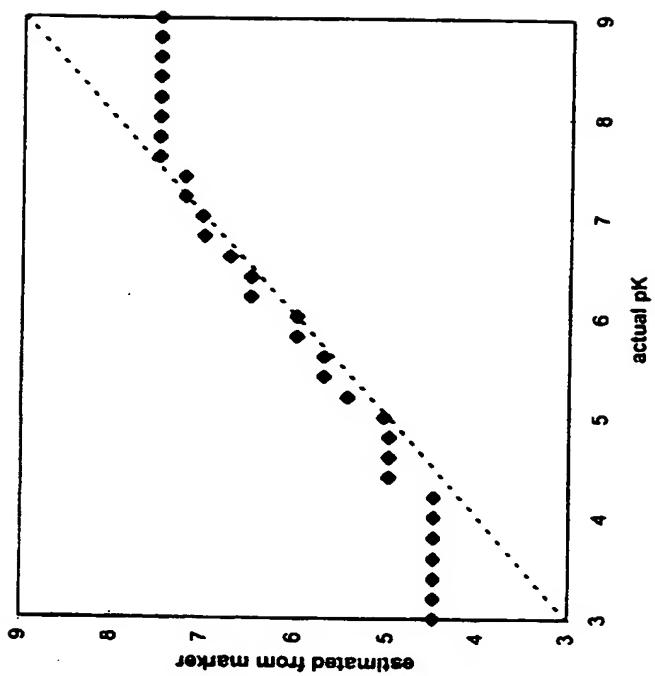


Fig. 16P

Quality of Estimation when Noise = 30 Inhibition Percentage Points

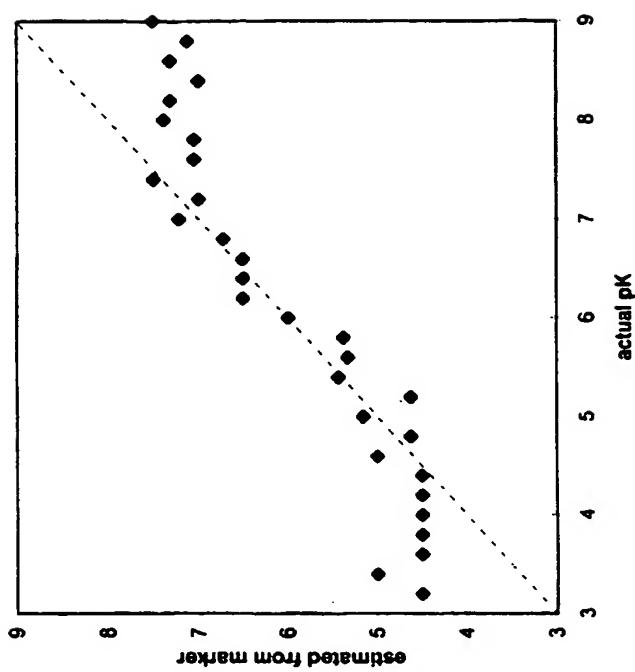
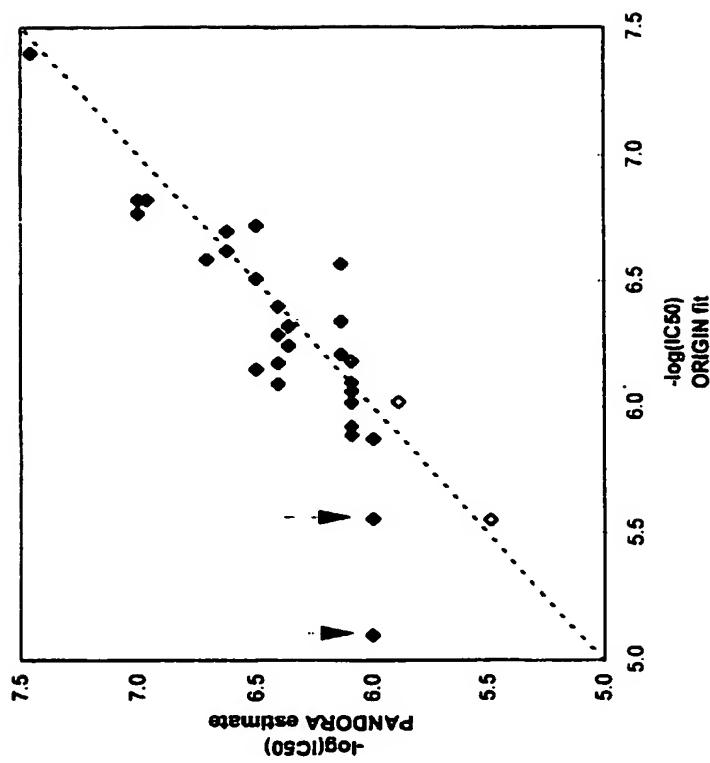


Fig. 16Q

**Comparison of Curve Fitting to Marker Calibration
for T-cell Proliferation Data**



	A	B	C	D	E	F	G	H	I
	Compd	Series	Test1	Test2	Test3	HTS SPA Dose-Resp % Inhib @ 3x10 ⁻⁶ M	HTS SPA Dose-Resp % Inhib @ 3x10 ⁻⁶ M	HTS SPA Dose-Resp % Inhib @ 3x10 ⁻⁷ M	HTS SPA Dose-Resp % Inhib @ 1x10 ⁻⁷ M
1									
2	Cmpd01	N	29	30	41	3	22	5	
3	Cmpd02	N	42	5.5	83	57	28	15	
4	Cmpd03	G	261	11	70	25	24	29	
5	Cmpd04	N		30	89	60	21	22	
6	Cmpd05	N	1.8	9.2	71	41	13	3	
7	Cmpd06	D	8.86	6.5	3.7	79	48	43	
8	Cmpd07	D	3.11	0.037	7.8	28	28	38	
9	Cmpd08	D		0.089	N.A.	65	22	15	
10	Cmpd09	O	0.119	0.233	68	41	24	5	
11	Cmpd10	N	4.31	4.31	61	42	63	25	
12	Cmpd11	N		50	77	24	3	3	
13	Cmpd12	H	1.3	0.24	47	25	40	37	
14	Cmpd13	H	1.17	0.194	30	59	4	12	
15	Cmpd14	H	0.26	0.41	39	23	46	36	
16	Cmpd15	H	0.369	0.148	99	46	38	18	
17	Cmpd16			0.87	30	101	47	24	
18	Cmpd17	K		0.223	N.A.	81	64	22	
19	Cmpd18		5.27			79	54	32	
20	Cmpd19		0.134			71	71	12	
21	Cmpd20			0.317		101	109	100	
22	Cmpd21			2.21		87	70	31	
23	Cmpd22	B		0.15		94	77	13	
24	Cmpd23	B				96	61	36	
25	Cmpd24	B		3.487	0.27	0.4	110	69	

Figure 17A

THIS COLORING INDICATES A DATA COLUMN WITH MIXED DATA TYPES										
orig col	heading	# numeric	# text	# date	# blank	# total	last	minimum	maximum	mean
A	Cmpd	24				24	25	(4 sig fig)	(4 sig fig)	(4 sig fig)
B	Series	20								
C	Test1	12				24	25	0.119	2.385	2.726
D	Test2	17				24	25	0.037	42	5.122
E	Test3	10	2			24	25	0.4	30	15.76
F	HTS SPA Dose-Resp % Inhib @ 3x10-8M	23				24	25	39	110	77.57
G	HTS SPA Dose-Resp % Inhib @ 1x10-6M	23				24	25	3	109	55.87
H	HTS SPA Dose-Resp % Inhib @ 3x10-7M	23				24	25	4	108	35.52
I	HTS SPA Dose-Resp % Inhib @ 1x10-7M	23				24	25	3	100	23.91
										(24 different)
										B{3} D{4} G{1} H{4} K{2} N{6}

Figure 17B

A	B	most important factor scored by Mngr A	most important factor scored by Mngr B	most important factor scored by Mngr C	less important factor scored by Mngr A	less important factor scored by Mngr B	less important factor scored by Mngr C	F	G
1	Proj 01	2							
2	Proj 02	1	2		1	2	2	2	
3	Proj 03		1		1	1	1	2	
4	Proj 04			1	1	2	1	1	
5	Proj 05				1	1	1	1	
6	Proj 06	2				2			
7	Proj 07				1	1	1	1	
8	Proj 08				1	1	1	1	
9	Proj 09				1	1	1	1	
10	Proj 10	2				2			
11	Proj 11	1				1			
12	Proj 12					2			
13	Proj 13								2
14	Proj 14					2			2
15	Proj 15	2				2		2	2
16	Proj 16	1	2		2	1	2	2	2
17	Proj 17	2				2			
18	Proj 18							1	1
19	Proj 19	2				2			
20	Proj 20					2			

Figure 18A

Click here to run these

sheet	Portfolio
column(s)	B:G
# of colors	3
break 1	1
break 2	2
break 3	3

color 1 red

color 2 yellow

color 3 black

Re-scale all?

Figure 18B

Name:	Factors
Enlarge Cluster Starts	
Sheet #	Portfolio
Cluster Col	A
Shrink Cluster Starts	
Color	Score
red	1
yellow	2
Score and Sort Clusters	
Column(s)	Rel. Weight
B:D	3
E:G	1

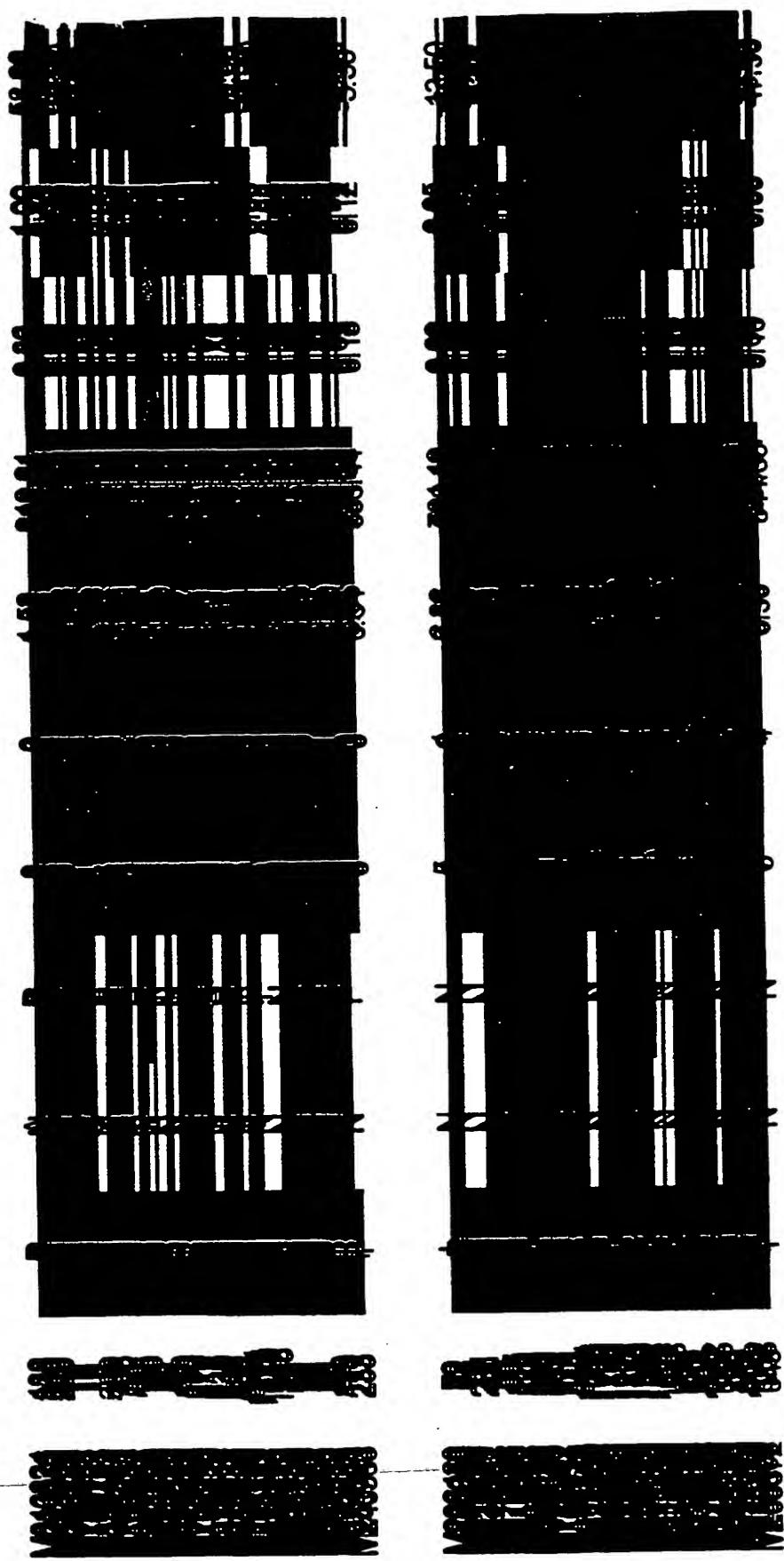
Figure 18C

	A	B	C	D	E	F	G	H
project name	most important factor scored by Mngr A	most important factor scored by Mngr B	most important factor scored by Mngr C	most important factor scored by Mngr A	less important factor scored by Mngr C	less important factor scored by Mngr B	less important factor scored by Mngr A	score (0-100)
1	Proj 01	18	2	2	2	1	2	94
2	Proj 02	05	04	14	2	2	1	92
3	Proj 03	05	04	14	2	2	1	86
4	Proj 04	05	04	14	2	2	1	83
5	Proj 05	05	04	14	2	2	1	83
6	Proj 06	05	04	14	2	2	1	81
7	Proj 07	05	04	13	2	2	1	75
8	Proj 08	05	04	13	2	2	1	75
9	Proj 09	05	04	13	2	2	1	72
10	Proj 10	05	04	13	2	2	1	69
11	Proj 11	05	04	12	2	2	1	67
12	Proj 12	05	04	12	2	2	1	67
13	Proj 13	05	04	12	2	2	1	64
14	Proj 14	05	04	12	2	2	1	61
15	Proj 15	05	04	11	2	2	1	58
16	Proj 16	05	04	11	2	2	1	58
17	Proj 17	05	04	10	2	2	1	58
18	Proj 18	05	04	09	2	2	1	39
19	Proj 19	05	04	09	2	2	1	36
20	Proj 20	05	04	08	2	2	1	2
21	Proj 21	05	04	08	2	2	1	1

Figure 18D

Fig. 19
Drug Candidate Compounds;
Scored, Sorted, and Vertically Compressed

Cmpd	Cluster	CD45	IRP	APOP	HTS	HTS	ACC	DON	clogP	MW	HTS	Selectivity	Ratio	Ratio	Pent	Cnt
------	---------	------	-----	------	-----	-----	-----	-----	-------	----	-----	-------------	-------	-------	------	-----



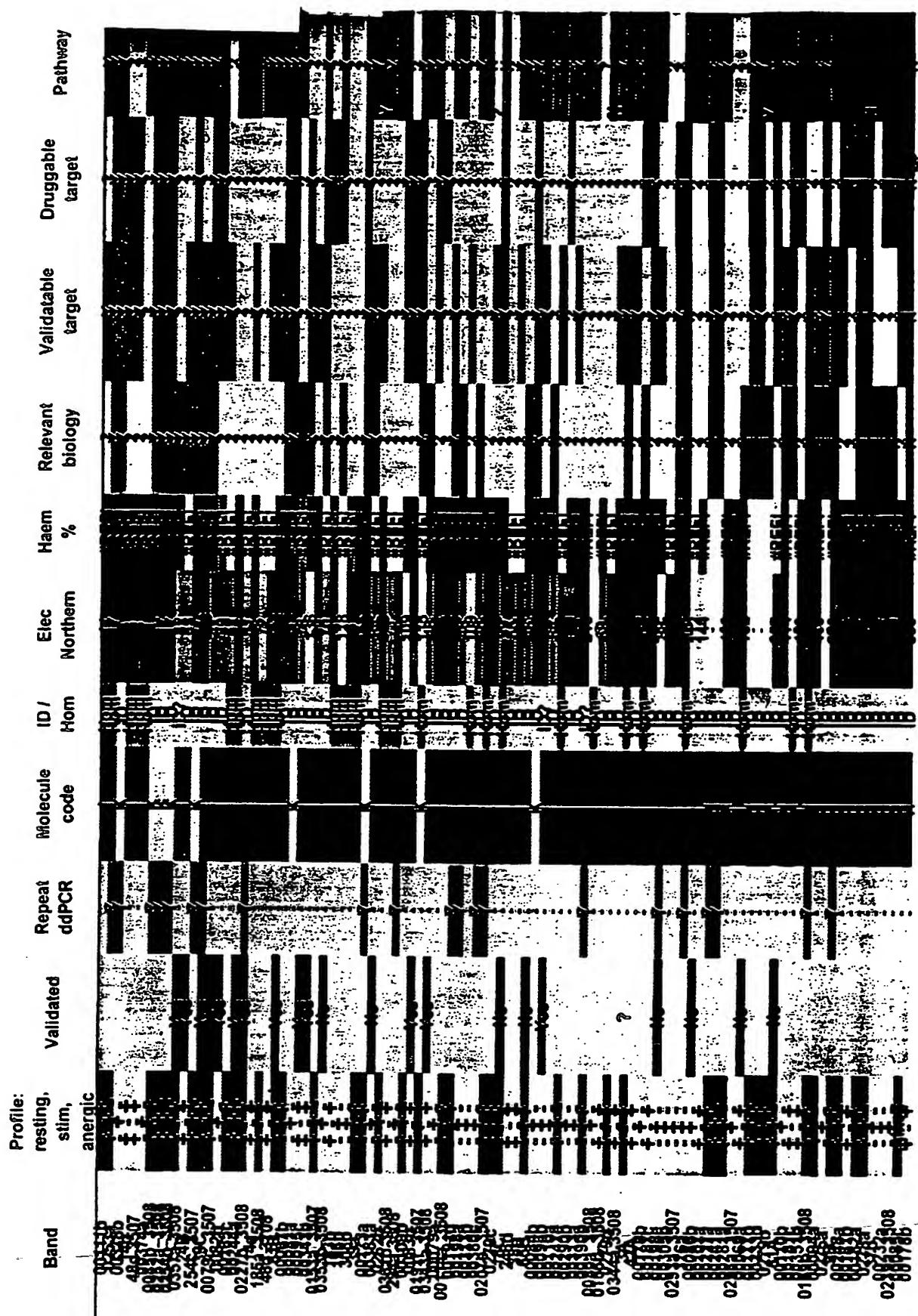


Fig. 20 Target Protein Candidates

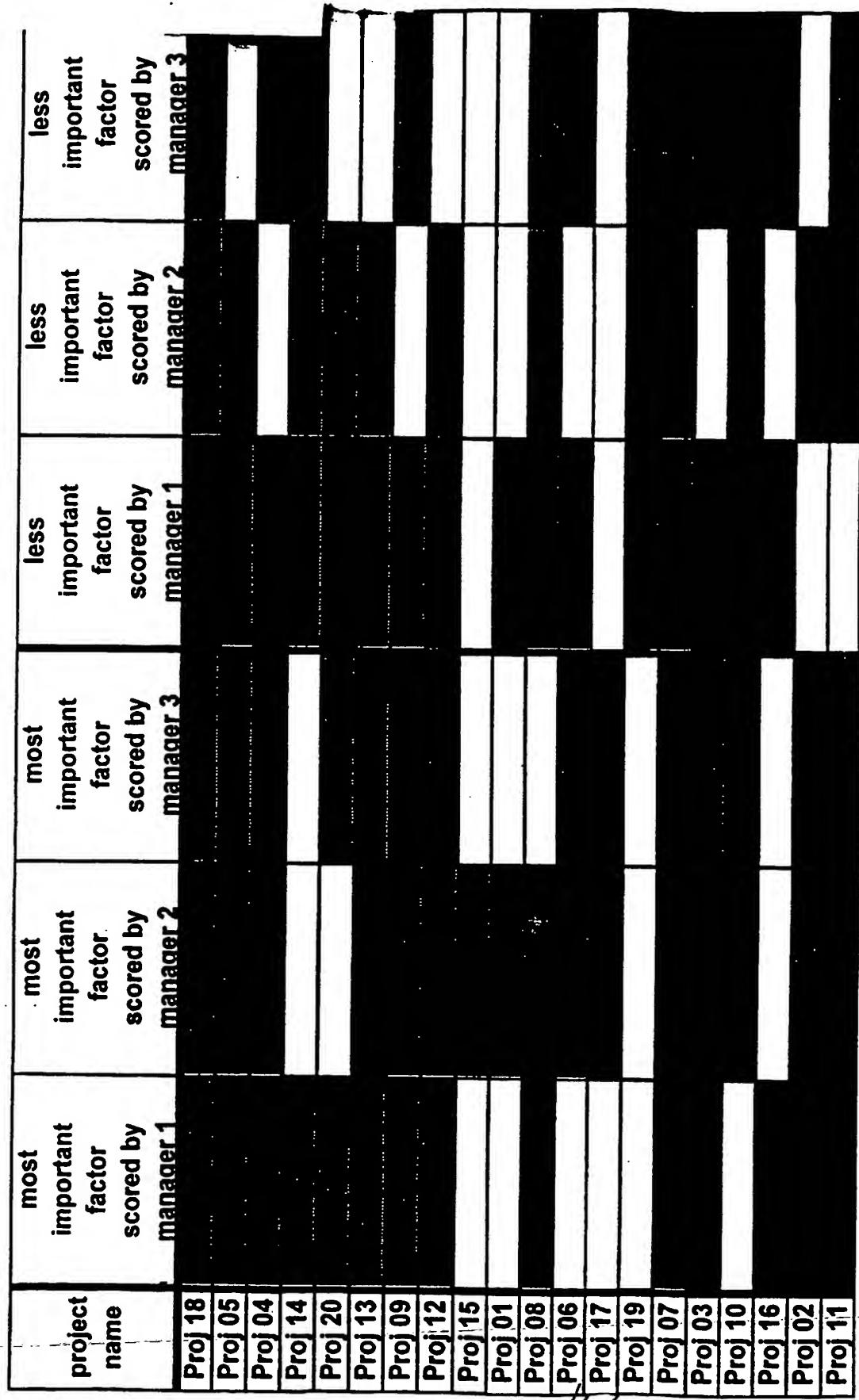


Fig. 21

Company	Disease 1	Disease 2	Disease 3	Disease 4	Disease 5	Disease 6	Disease 7	Disease 8	Disease 9	Disease 10	Disease 11	Disease 12	Disease 13
Company 1	Phase I		Phase II		LO/DE		LO/DE		LO/DE	Phase I	Phase II	Phase II	Phase II
Company 2	Phase I	LI	Phase II	LI	LI/LO	LI/LO	LI/LO	LI/LO	LI/LO	LI/LO	LI/LO	LI/LO	LI/LO
Company 3	LI/LO				LI/LO	LI/LO	LI/LO	LI/LO	LI/LO	LI/LO	LI/LO	LI/LO	LI/LO
Company 4	LO/DE				LO/DE	LO/DE	LO/DE	LO/DE	LO/DE	LO/DE	LO/DE	LO/DE	LO/DE
Company 5	LI/LO				LO/DE	LO/DE	LO/DE	LO/DE	LO/DE	LO/DE	LO/DE	LO/DE	LO/DE
Company 6	LO/DE				LO/DE	LO/DE	LO/DE	LO/DE	LO/DE	LO/DE	LO/DE	LO/DE	LO/DE
Company 7	LI/LO				LO/DE	LO/DE	LO/DE	LO/DE	LO/DE	LO/DE	LO/DE	LO/DE	LO/DE
Company 8	LO/DE	Phase II			LO/DE	LO/DE	LO/DE	LO/DE	LO/DE	LO/DE	LO/DE	LO/DE	LO/DE
Company 9	LI/LO				LO/DE	LO/DE	LO/DE	LO/DE	LO/DE	LO/DE	LO/DE	LO/DE	LO/DE
Company 10					LO/DE	LO/DE	LO/DE	LO/DE	LO/DE	LO/DE	LO/DE	LO/DE	LO/DE
Company 11	LO/DE				LO/DE	LO/DE	LO/DE	LO/DE	LO/DE	LO/DE	LO/DE	LO/DE	LO/DE
Company 12	LO/DE				LO/DE	LO/DE	LO/DE	LO/DE	LO/DE	LO/DE	LO/DE	LO/DE	LO/DE
Company 13	LI				LI/LO	LI/LO	LI/LO	LI/LO	LI/LO	LI/LO	LI/LO	LI/LO	LI/LO
Company 14	LI/LO				LI	LI	LI	LI	LI	LI	LI	LI	LI
Company 15	LI				LI	LI	LI	LI	LI	LI	LI	LI	LI
Company 16	LI				LI	LI	LI	LI	LI	LI	LI	LI	LI
Company 17	Phase II				Phase II	Phase II	Phase II	Phase II					
Company 18					LI	LI	LI	LI	LI	LI	LI	LI	LI
Company 19	LI				LI/LO	LI/LO	LI/LO	LI/LO	LI/LO	LI/LO	LI/LO	LI/LO	LI/LO
Company 20					TS	TS	TS	TS	TS	TS	TS	TS	TS
Company 21	LI/LO				TS	TS	TS	TS	TS	TS	TS	TS	TS
Company 22	TS				TS	TS	TS	TS	TS	TS	TS	TS	TS
Company 23	LI				TS	TS	TS	TS	TS	TS	TS	TS	TS
Company 24	LI				TS	TS	TS	TS	TS	TS	TS	TS	TS
Company 25	LI/LO				TS	TS	TS	TS	TS	TS	TS	TS	TS

Fig. 22



Fig. 23

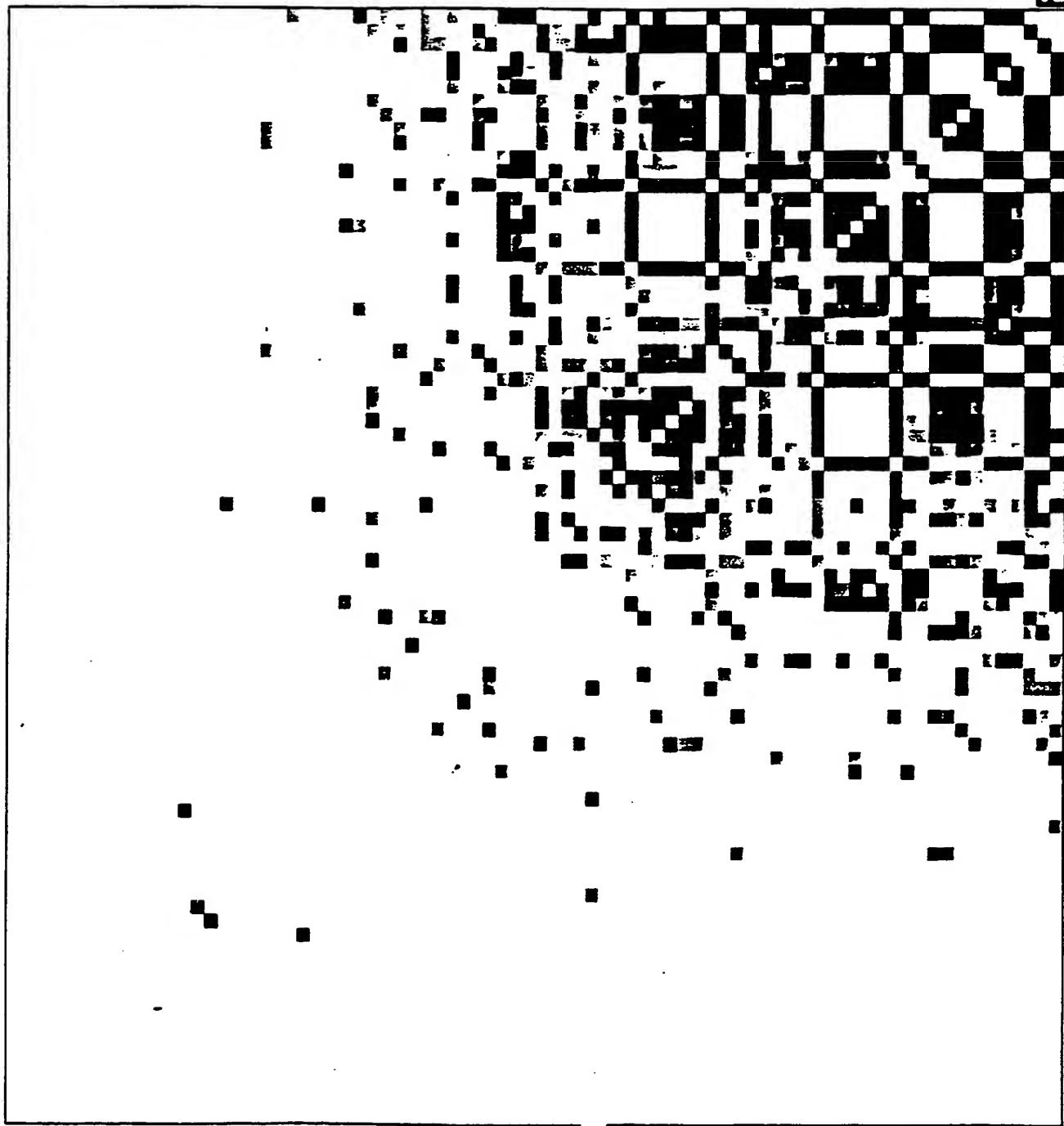
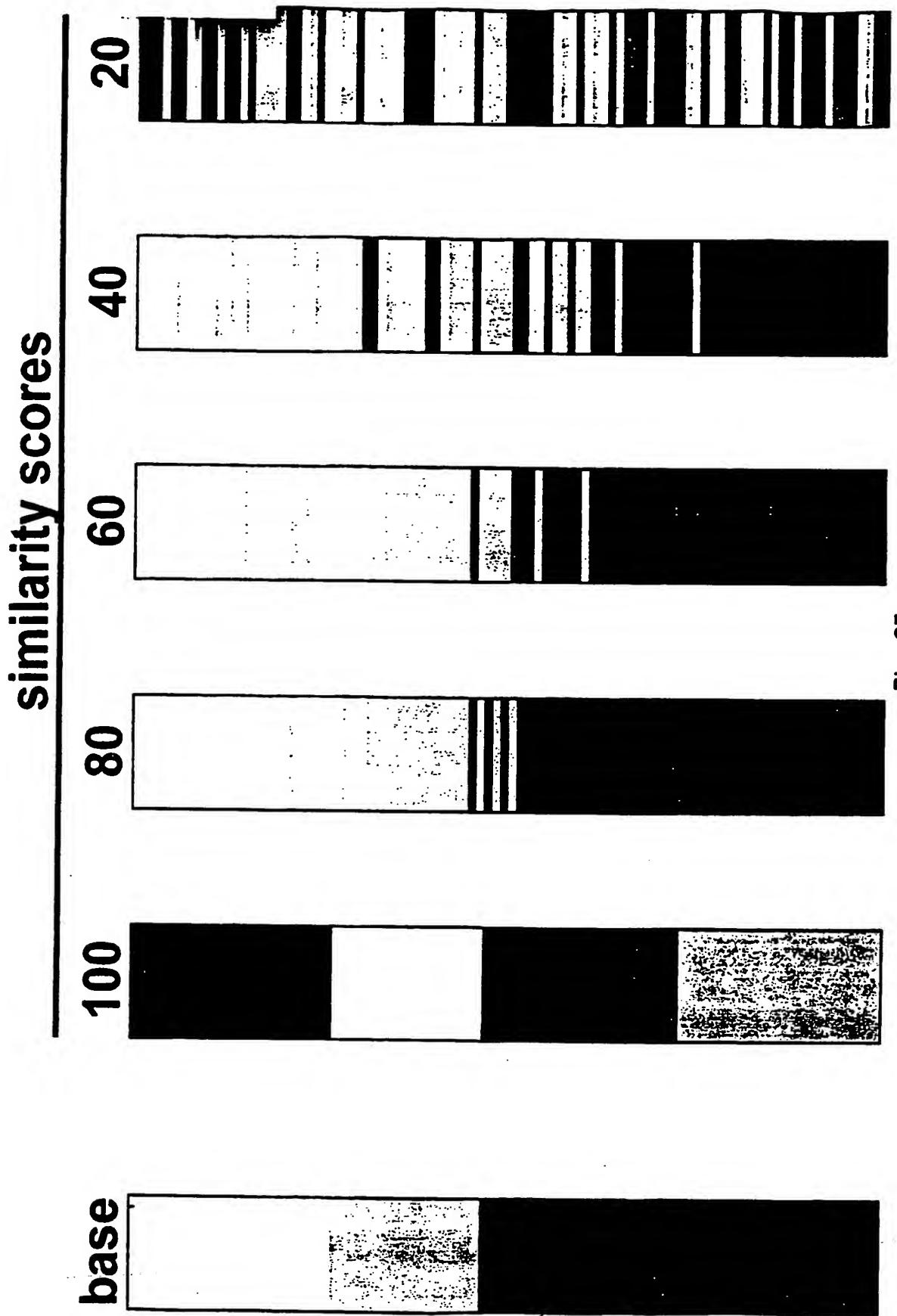


Fig. 24



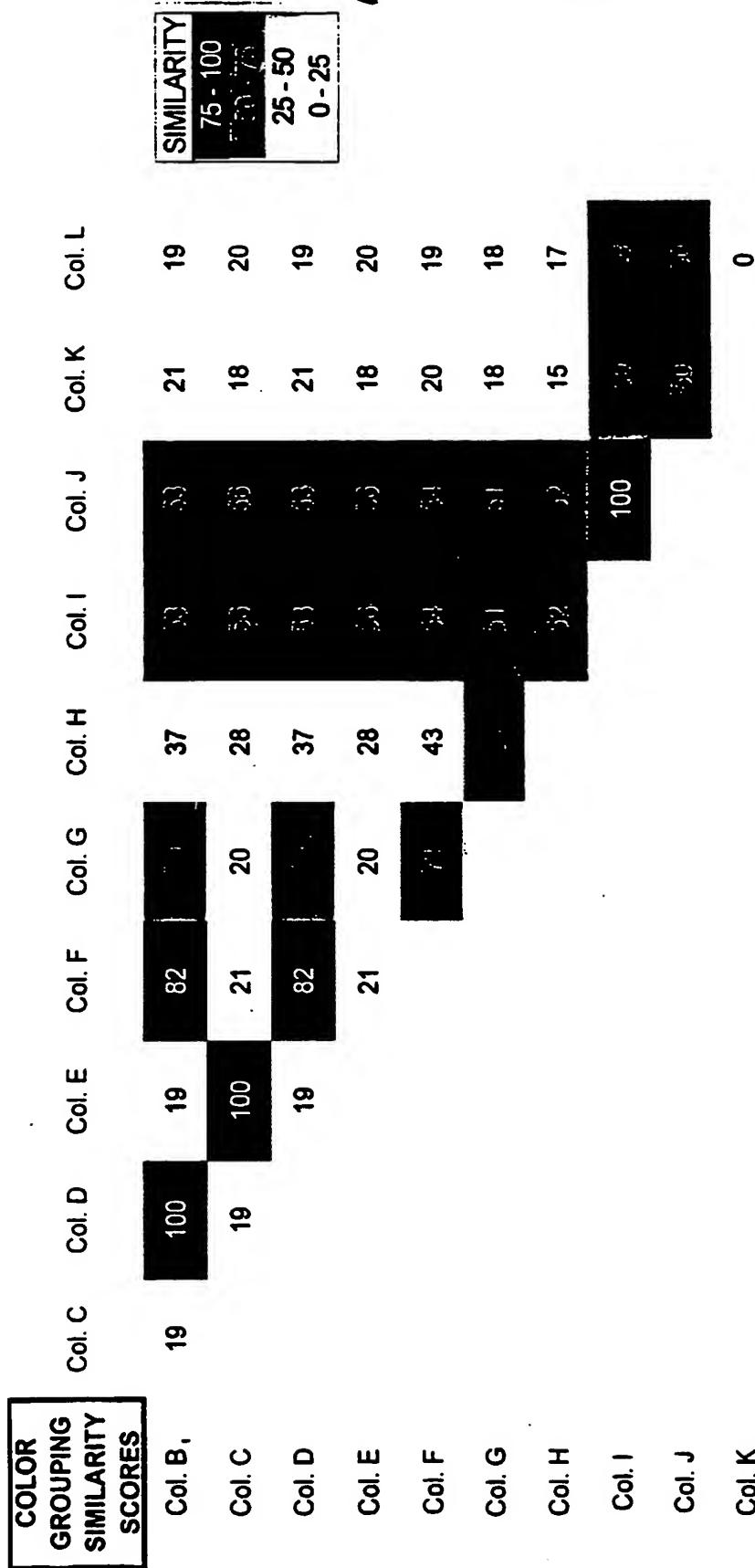


Fig. 26

(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization
International Bureau



(43) International Publication Date
1 February 2001 (01.02.2001)

PCT

(10) International Publication Number
WO 01/08039 A3

(51) International Patent Classification⁷: **G06K 9/00**

(21) International Application Number: **PCT/US00/20401**

(22) International Filing Date: 27 July 2000 (27.07.2000)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
09/361,122 27 July 1999 (27.07.1999) US

(71) Applicant (for all designated States except US): **ZENECA LIMITED** [GB/GB]; 15 Stanhope Gate, London W1Y 6LN (GB).

(72) Inventor; and

(75) Inventor/Applicant (for US only): **LERMAN, Charles, L.** [US/US]; 501 Bishop Hollow Road, Newton Square, PA 19073-3138 (US).

(74) Agents: **BIRD, Donald, J. et al.**; Pillsbury Madison & Sutro, LLP, 1100 New York Avenue, N.W., Washington, DC 20005 (US).

(81) Designated States (national): AE, AG, AL, AM, AT, AU, AZ, BA, BB, BG, BR, BY, BZ, CA, CH, CN, CR, CU, CZ, DE, DK, DM, DZ, EE, ES, FI, GB, GD, GE, GH, GM, HR, HU, ID, IL, IN, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MA, MD, MG, MK, MN, MW, MX, MZ, NO, NZ, PL, PT, RO, RU, SD, SE, SG, SI, SK, SL, TJ, TM, TR, TT, TZ, UA, UG, US, UZ, VN, YU, ZA, ZW.

(84) Designated States (regional): ARIPO patent (GH, GM, KE, LS, MW, MZ, SD, SL, SZ, TZ, UG, ZW), Eurasian patent (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European patent (AT, BE, CH, CY, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE), OAPI patent (BF, BJ, CF, CG, CI, CM, GA, GN, GW, ML, MR, NE, SN, TD, TG).

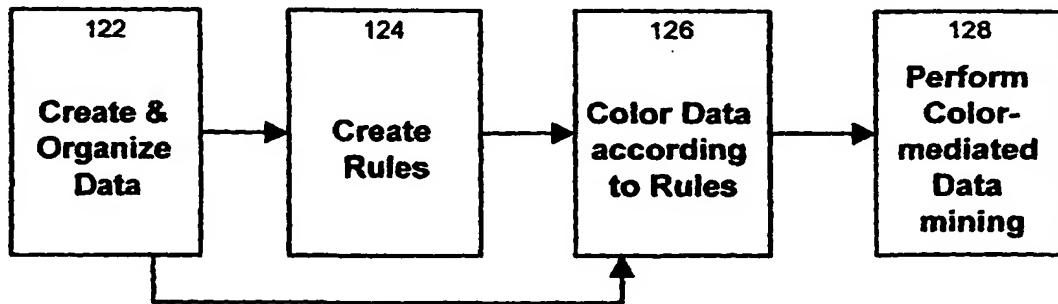
Published:

- With international search report.
- Before the expiration of the time limit for amending the claims and to be republished in the event of receipt of amendments.

(88) Date of publication of the international search report: 22 March 2001

For two-letter codes and other abbreviations, refer to the "Guidance Notes on Codes and Abbreviations" appearing at the beginning of each regular issue of the PCT Gazette.

(54) Title: ANALYSIS AND PATTERN RECOGNITION IN LARGE, MULTIDIMENSIONAL DATA SETS USING LOW-RESOLUTION DATA GROUPING



(57) Abstract: Methods, systems and devices for operating on data provide at least one user-defined grouping rule for grouping the data into a user-definable number of groups; and apply at least one of the grouping rules to the data. The data may be in a table, wherein the at least one grouping rule applies to at least one user-selectable column of the table. The grouping rule defines breakpoints corresponding to the user-definable number of groups, and application of the at least one rule to the data divides the data into groups based on the breakpoints. The grouped data is presented in a manner that visually distinguishes the groups, sometimes by coloring an aspect of the data according to the rules.

WO 01/08039 A3

INTERNATIONAL SEARCH REPORT

International Application No
PCT/US 00/20401

A. CLASSIFICATION OF SUBJECT MATTER

IPC 7 G06K9/00

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

IPC 7 G06K

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practical, search terms used)

EPO-Internal, COMPENDEX, INSPEC, WPI Data, IBM-TDB, PAJ

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category *	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	<p>ANONYMOUS: "Dynamic Layout Mechanism for the Massive-Node Server Status Monitor" IBM TECHNICAL DISCLOSURE BULLETIN, vol. 36, no. 5, 1 May 1993 (1993-05-01), pages 169-170, XP000408951 New York, US the whole document</p> <p style="text-align: center;">-/-</p>	1-48

Further documents are listed in the continuation of box C.

Patent family members are listed in annex.

* Special categories of cited documents :

- *A* document defining the general state of the art which is not considered to be of particular relevance
- *E* earlier document but published on or after the international filing date
- *L* document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)
- *O* document referring to an oral disclosure, use, exhibition or other means
- *P* document published prior to the international filing date but later than the priority date claimed

- *T* later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention
- *X* document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone
- *Y* document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art.
- *&* document member of the same patent family

Date of the actual completion of the international search

12 January 2001

Date of mailing of the international search report

22/01/2001

Name and mailing address of the ISA

European Patent Office, P.B. 5818 Patentlaan 2
 NL - 2280 HV Rijswijk
 Tel. (+31-70) 340-2040, Tx. 31 651 epo nl.
 Fax (+31-70) 340-3016

Authorized officer

Granger, B

INTERNATIONAL SEARCH REPORT

Inter [REDACTED] Application No
PCT/US 00/20401

C.(Continuation) DOCUMENTS CONSIDERED TO BE RELEVANT

Category	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	<p>EISEN M B ET AL: "Cluster analysis and display of genome-wide expression patterns" PROCEEDINGS OF THE NATIONAL ACADEMY OF SCIENCES OF USA, US, NATIONAL ACADEMY OF SCIENCE, WASHINGTON, vol. 95, December 1998 (1998-12), pages 14863-14868, XP002140966 ISSN: 0027-8424 page 14863, right-hand column, paragraph 3; figure 2</p> <p>-----</p> <p>STANTON D T ET AL: "Application of nearest-neighbor and cluster analyses in pharmaceutical lead discovery" JOURNAL OF CHEMICAL INFORMATION AND COMPUTER SCIENCES, JAN.-FEB. 1999, ACS, USA, vol. 39, no. 1, pages 21-27, XP000971515 ISSN: 0095-2338 the whole document</p> <p>-----</p>	1-48
A		1-48